CFPS China Family Panel Studies

China Family Panel Studies (2010)

User's Manual

Yu Xie

12/20/2012

CONTENTS

FOREWORD **1**

ACKNOWLEDGEMENT **3**

1. INTRODUCTION **5**

1.1 Background and Baseline Survey 5
1.2 Design 6
1.3. International Comparison 12
1.4. Survey Technologies 14

2. SAMPLING **16**

2.1 Sampling Design 16
2.2 Terminal Sampling Frame 19

3. QUESTIONNAIRE DESIGN **21**

3.1 Overview 21
3.2 Community questionnaire 24
3.3. Residence Screening 25
3.4 Household Screening Questionnaire 25
3.5 Family roster questionnaire 26
3.6 Family Questionnaire 31
3.7 Individual Questionnaire 33

4. IMPLEMENTATION **37**

4.1 Pilot studies 37
4.2 Interviewers 37
4.3 Overview of 2010 Survey Implementation 38
4.4 Refusal and solutions 39
4.5 Baseline survey final contact result 40
4.6 Sample maintenance 43

5. SURVEY QUALITY CONTROL **44**

5.1 Quality Control Measures and Technologies 44
5.2 Quality Control Strategies 45
5.3 Proportions and Results of Quality Check 45

6. DATA SETS AND VARIABLES **47**

6.1 General Introduction to Data Sets 47
6.2 Data Cleaning 47
6.3 Composite Variables 52
6.4 Occupation Codes 61
6.5 Confidentiality Issues 63
6.6 Miscellaneous 63

7. CFPS 2010 BASELINE SURVEY PRELIMINARY FINDINGS AND EVALUATIONS **66**

7.1 AGE-SEX DISTRIBUTIONS 66
7.2 FAMILY SIZE AND HOUSEHOLD TYPES 70
7.3 FAMILY INCOME 71
7.4 URBAN-RURAL DISTRIBUTION 74
7.5 EDUCATIONAL LEVEL 75
7.6 MARITAL STATUS 76

8. WEIGHTS CALCULATIONS 78

9. TECHNICAL REPORTS 80

10. REFERENCES 82

Foreword

After several years of preparation and two pilot studies in 2008 and 2009, China Family Panel Studies (CFPS) implemented its baseline survey in 2010. Survey research is known to be complicated and involves a multitude of details—including those in the initial conceptual design, survey technologies, interviewing processes, quality control, and the construction of data sets and data cleaning. Each individual aspect has influences on the academic value of the data. The CFPS sample covers 25 provinces/municipalities/autonomous regions, representing 95% of the Chinese population. The 2010 baseline survey interviewed a total of 14,960 household and 42,590 individuals, who will participate in a long-term follow-up survey. The survey is China's first large-scale academic-oriented panel survey project. It aims to become the most authoritative survey project on Chinese family and society. While we are delighted to have accomplished the collection of comprehensive, high-quality, and valuable social science data, it comes with a cost—due to the complexity in the design, implementation, and the data structure, users might experience difficulty in using the data. Therefore, we publish this User's Manual to provide detailed information for the usage of the 2010 data in a systematic and user-friendly way. Specifically, we cover the following topics:

1. Conceptual design and methods, including those pertaining to sampling, weighting, survey instrument, questionnaire design, and follow-up strategies, etc.
2. Details of the actual implementation, such as map drawing, residence and household screening, standardization of interviewing procedures, data quality control, sample maintenance, etc.
3. Data management and data sets, including the structure and content of the data sets, data cleaning, construction of composite measures, occupational codes, etc.
4. Technical report index. Regarding certain substantive topics and professional fields, we provide a series of separate technical reports for users to better understand our project and data. This manual references and cites some parts of the technical reports, but does not elaborate on the content covered by the technical reports. Instead, an index of the reports is provided toward the end.
5. Data quality assessment. Based on comparisons with Chinese censuses and other data, we provide a brief evaluation on the quality of the CFPS 2010 baseline survey.

This manual draws extensively on the conference minutes, official documents, manuals, and technical reports provided by the staff and researchers from Institute of Social Science Survey (ISSS) at Peking University. The tables and graphs in the preliminary findings and evaluations were prepared by Chunni Zhang, Qi Xu, Xiang Zhou, Hongwei Xu, and Guoying Huang. The manual was organized and edited by

Jingwei Hu and proofread by Chunni Zhang. In addition, Yan Sun, Qiong Wu, and Ping Lv assisted with the writing of this manual.

We sincerely hope that data users find this manual helpful. If we need to make changes to the manual due to data or documentation updates, we will continue to publish updated versions. Please let us know if you identify any problems with this manual. Suggestions for improvement are also welcome. We will make sure to consider your ideas in our preparation for new versions.

Acknowledgement

A large number of people have made great and selfless contributions to the CFPS 2010 baseline survey. Our accomplishments in 2010 are a result of collective efforts. Here we would like to acknowledge and express our appreciation for the entire CFPS team.

Led by Peking University, many researchers and scholars from home and abroad joined the design of CFPS baseline survey questionnaires. The main contributors include Yude Chen, Yuyu Chen, Xiaohao Ding, Jiafeng Gu, Zhigang Guo, Guitian Huang, Guoping Li, Jianxin Li, Shiding Liu, Xiaochun Qiao, Zeqi Qiu, Mingming Shen, Yan Sun, Ping Tu, Qun Xiao, Yu Xie, Xianglin Xu, Jie Yan, Boxu Yang, Yang Yao, Ruijun Yuan, Changjun Yue, Qianfan Zhang, Tuohong Zhang, Yaohui Zhao, Xiaolin Zhou, Yanhui Zou, He Cai (Sun Yat-Sen University), Youde Guo (Fudan University), Hong Lei (Huazhong University of Science and Technology), Lulu Li (Renmin University of China), Peilin Li (Chinese Social Science Academy), Shi Li (Beijing Normal University), Youmei Li (Shanghai University), Jingming Liu (Tsinghua University), Yuzhao Liu (Shanghai University), Liping Qiu (Shanghai University), Guangzhou Wang (Chinese Social Science Academy), Dingjun Weng (Shanghai University), Xiaogang Wu (Hong Kong University of Science and Technology), Jianhua Zhu (City University of Hong Kong), Colette Browning (Monash University), Michael Carter (University of Wisconsin-Madison), RobertHauser (National Research Council and the University of Wisconsin-Madison), David Lam (University of Michigan), James Lepkowski (University of Michigan), Arland Thornton (University of Michigan), Donald Treiman (University of California-Los Angeles), Nora Schaeffer (University of Wisconsin-Madison), Robert Willis (University of Michigan), Jean Yeung (National University of Singapore). CFPS questionnaires cover almost all fields of social science research. We would like to thank the above researchers and scholars for their constructive opinions and suggestions, which were invaluable for enriching, improving, and optimizing the content of the questionnaires.

At the same time, we also thank all those who had contributed to the survey implementation. During the busy seasons of survey, they often work overtime, even on holidays, to effectively deal with all kinds of problems that emerged from the fields. We owe the successful implementation to the hard work by He Cai, Yulong Cao, Minyan Chen, Wei Cong, Jiapo Chen, Lijuan Ci, Hua Ding, Chunjie Gu, Jiafeng Gu, Zhenwei Guo, Junli Han, Changqun Huang, Danli Jia, Guohua Li, Li Li, Shengwen Li, Youmei Li, Yucheng Liang, Yue Liu, Ping Lv, Ping Qian, Zeqi Qiu, Liyin Ren, Shibin Song, Shuai Sun, Yan Sun, Yi Sun, Yuhuan Sun, Xueliang Teng, Xu Yang, Rong Shen, Tao Wang, Yanmei Wang, Qi Xu, Jie Yan, Qian Yang, Jiahui Yao, Xue Ye, Wenmao Yin, Shuang Yu, Xuejun Yu, Haobin Zang, Haidong Zhang, Lanxin Zhang, Man Zhang, Yaxin Zhang, Yun Zhou, Tingwei Zhu, Yanhui Zou.

Due to the large sample size and multiple types of data, the management and cleaning of CFPS data have involved a great amount of workforce. The CFPS data team has worked extremely hard to provide reliable and user-friendly data, for which we would like to thank Ling Bai, Hua Ding, Jingwei Hu, Guoying Huang, Li Li, Wangyang Li, Weixiang Luo, Ping Lv, Sha Ni, Liyin Ren, Qianping Ren, Qiang Ren, Yuhuan Sun, Zhibo Tan, Jia Wang, Xiao Wang, Lingwei Wu, Qiong Wu, Jun Xiang, Yu Xie, Hongwei Xu, Qi Xu, Jie Yan, Jiahui Yao, Jia Yu , Chunni Zhang, Duan Zhao.

The implementation of CFPS has been assisted by the former National Population and Family Planning Commission, State Bureau of Statistics, Shanghai University, and Sun Yat-Sen University. The Institute for Social Research at the University of Michigan, our important collaborative unit, has provided us with much guidance and advice on survey design and technological support. In addition, our survey is supported by 985 Project funds at Peking University, to which we are deeply grateful.

Finally and most importantly, we would like to thank the more than 400 interviewers who conducted interviews in various study sites across the country in 2010. Fieldwork is the most important and yet the most laborious part of survey research. Our interviewers have overcome numerous transportation difficulties and achieved excellence in their work, and offered their blood, toil, tears, and sweat in exchange for the high quality data. In addition, we are indebted to our more than 40,000 respondents, whose understanding and cooperation made our study possible. Without our respondents, it would not have been possible to acquire such precious empirical materials that truthfully reflect China's social conditions.

1. Introduction

1.1 Background and Baseline Survey

China Family Panel Studies (CFPS) is a national longitudinal general social survey project. By collecting data at three levels (i.e., individual, family, community), the project aims to document changes in Chinese society, economy, population, education, and health, so as to provide data for academic research and public policy analysis.

CFPS focuses on economic and non-economic well-being of the Chinese people, covering substantive areas such as economic activities, educational attainment, family relationships and dynamics, migration, and physical and mental health. The target sample of CFPS consists of 16,000 households in 25 provinces/municipalities/autonomous regions in China (excluding Hong Kong, Macao, Taiwan, Xinjiang, Tibet, Qinghai, Inner Mongolia, Ningxia and Hainan). All eligible households and household members are subjects of the survey. An eligible household refers to an independent economic unit that lives in a residential community and one or more family members are of Chinese nationality.[1] Family members are defined as finaicially dependent immediate relatives[2], or non-immediate blood/marital/adoptive relatives who have lived with the household for more than three consecutive months and financially related to the sampled household.

The preparatory work of CFPS started in 2007. Pilot studies were conducted with 2,400 households in Beijing, Shanghai and Guangdong, including a baseline survey in 2008 and a follow-up survey in 2009. In 2010, the baseline national survey was officially launched in 25 provinces/municipalities/autonomous regions. We visited 19,986 households and successfully interviewed 14,960 households. Within these households, a total of 33,600 adults and 8,990 youths were interviewed. At the household level, the response rate, cooperation rate, contact rate, refusal rate are 81.25%, 96.58%, 84.13%, 2.67%, respectively. At the individual level, the rates are 84.14%, 87.01%, 96.7%, and 8.47%, respectively. [3]

The CFPS 2010 baseline survey conducted face-to-face interviews with the sampled households' family members who live in the sample communities. It also interviewed those family members who were elsewhere in the same county. For those who were not present at home at the time of interview, basic information was collected from their family members at presence. All family members who had blood/marital/adoptive ties with the household were identified as permanent respondents. Prospective family members including new-borns and adopted children

[1] Initially at least one family member who lives in our sample community for consecutive six months is required for household eligibility. However, the requirement is cancelled during survey implementation since most of our sampled households fulfill this requirement.
[2] See Sun et al. (2011) for definition of immediate family members
[3] Based on AAPOR estimation. See Technical Report: CFPS-5.

in the future will also be considered as permanent respondents. All permanent respondents are to be tracked in all follow-up surveys.

CFPS 2010 baseline survey adopted face-to-face interview aided by Computer Assisted Personal Interviewing (CAPI) technology. Such design has enabled diversified questionnaire designs, high interviewing efficiency, and good data quality. CFPS 2010 used five questionnaires targeted at community, family members, households, adults, and youths. These five questionnaires founded our base for future questionnaire designs aiming at different respondants and interview techniques, such as phone interviews and online interviews.

CFPS is designed by Peking University research team and supported by Peking University 985 funds. The Institute of Social Science Survey (ISSS) at Peking University is responsible for the implementation, which has received great support from former Chinese National Population and Family Planning Commission.

1.2 Design

1.2.1 Social Changes in China[4]

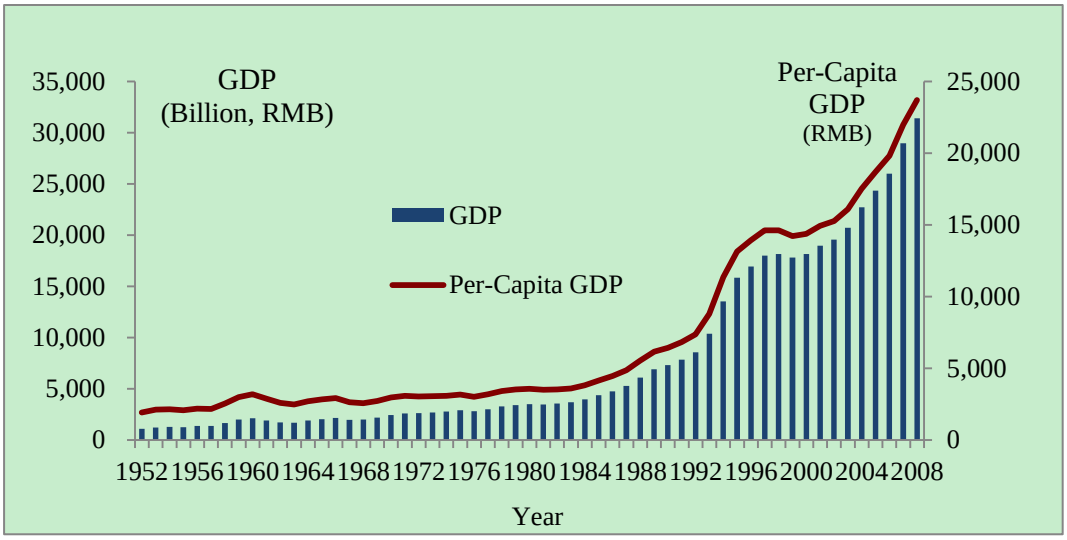
China has been undergoing a social transformation whose scope, rapidity, and influences are unprecedented in human history. In a sense, China's ongoing social transformation since the late twentieth century is no less consequential in the course of world history than events commonly considered as historical watersheds, such as the Italian Renaissance, the Protestant Reformation, or the British Industrial Revolution. The rapid, large-scale, and irreversible social changes that have been occurring in China in recent decades are multifaceted. We may gain a better appreciation of those social changes by looking at three highly important aspects—economic growth, education expansion, and demographic transition.

Chinese economic output grew tremendously and sustanablly since the 1980s. Figure 1 shows a sharp increase in total gross domestic product (GDP) and per-capita GDP began after the initiation of economic reform in 1978. The per-capita inflation-adjusted GDP increased at an annual growth rate of 6.7 % between 1978 and 2008. Compared to the massive, sustained, and rapid economic development in China, the 1.5 % annual growth rate in inflation-adjusted per-capita GDP during the Gilded Age in United States was much smaller.[5]

In addition, educational attainment in China has been significantly improved in recent years, most dramatically at the postsecondary level. Figure 2 shows trends in the numbers of enrolled and newly admitted college students. While the figure shows a gradual increase in college enrollments over time—except for a downturn associated

[4] Sections 1.2.1、1.2.2、1.2.3 draw partially on Xie (2011).
[5] Calculation based on data from Measuring Worth (2011) http://www.measuringworth.com/.

with the Cultural Revolution (1966–1976)—the boost began in the late 1990s. The rapid increase in the number of young Chinese receiving higher education is both cause and consequence of China's tremendous economic growth in recent decades.



Note: Adjustment has been done for the data of 2005–2008, on the basis of the 2d Economic Census.
Sources: State Bureau of Statistics. 2010. *China Statistical Abstracts 1949-2008*. China Statistics Press. State Bureau of Statistics. 2010. *China Statistical Abstracts 1949-2008*. China Statistics Press.

Figure 1. Trends in GDP and Per-Capita GDP, 1952–2008 (in 2008 RMB).

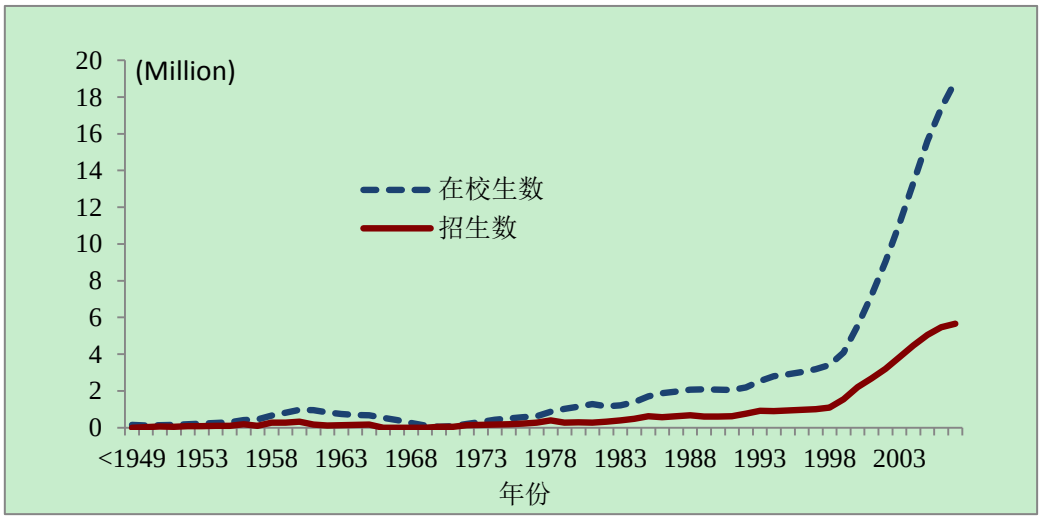
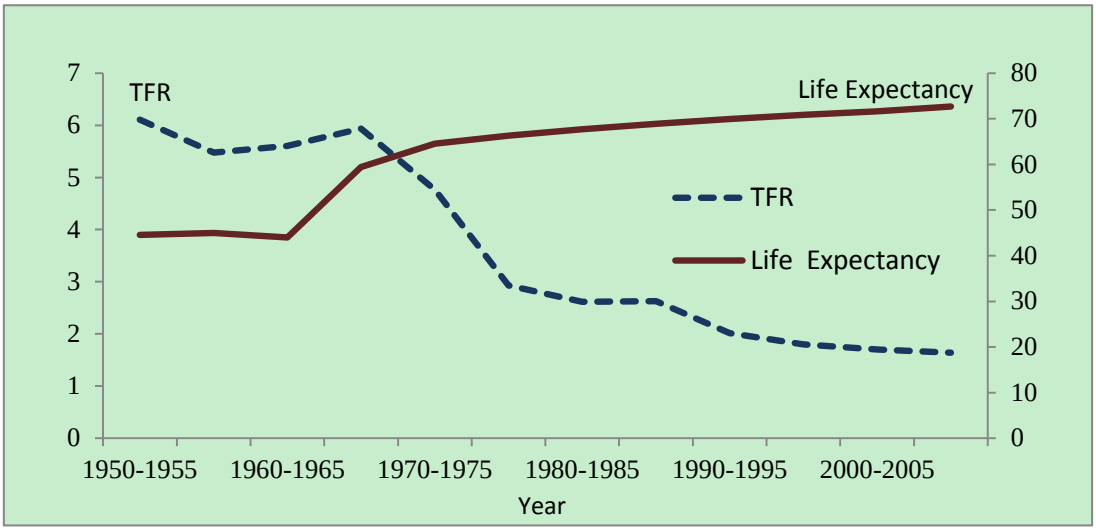


Figure 2. Higher Education Expansion in China, 1949–2007. Source: China Education Yearbook Editorial Board (1984, 1986–1988, 1989–2008).

Furthermore, China has completed its demographic transition, from a high-fertility, high-mortality regime to a low-fertility, low-mortality regime during the past decades. Clearly in Figure 3, there is a sharp decline in total fertility rate (TFR) since

the late 1970s, as well as a steady improvement in life expectancy since the 1950s. By 1990–1995, China had already completed its demographic transition, with TFR decreased to the replacement level at 2.0, and life expectancy increased to a comparable level in developed countries around 1970 at 70 while way ahead of other underdeveloped.



Sources: United Nations, Department of Economic and Social Affairs, Population Division (2011). World Population Prospects: The 2010 Revision, CD-ROM Edition.

Figure 3. Total Fertility Rate and Life Expectancy, 1950-2010.

The aforementioned social changes are probably the most significant ones in China over the past thirty to forty years. Of course, many other important social changes have taken place which call for social science research. Examples include increasing social inequality, rising divorce and premarital cohabitation, and massive labor migration. It is CFPS' goal to provide social scientists with high-quality, comprehensive, and longitudinal data to conduct in-depth research on such social phenomena and their changes.

1.2.2. The Importance of Empirical Research in China

The research in studying the uniqueness of the social changes happened in China does not only provide valuable findings, but also has the potential to change the course of world history. For the last three centuries or so, the West has been leading the world as its ascendant path has commonly been used to describe "modernization," "development," and "progress." The foundations of the Western countries' development are (1) democracy as the only legitimate political system, and (2) the free market as the only viable economic system. For the first time in the past three centuries, China is now presenting a serious challenge to the West as the modernization theory fails to explain the development path of China. China neither becomes a democratic country, nor establishes a truly free market economic system.

However, its economy has been developing rapidly and steadily for more than 30 years. In contrast, western economies, such as that in the U.S., have been stagnating and faltering in recent years. Is it possible that China's model may also be a feasible, perhaps even superior, path to development?

The China's development model is an interesting topic which draws our immediate attention. The study on such social science issues requires objective judgement without prejudice, which is often grounded in projections of experiences from other countries or from theoretical speculations. We may understand China's situation from its history, culture, politics, and economy, and analyze based on empirical evidence rather than pure speculation.

Social theory is best based on a social context, and social research is best conducted with a specific social context. Thus the large-scale rapid social changes in contemporary China are especially worth studying with its unique and exclusive characteristics among exsiting and perspective social reforms from ancient to modern times. China's current development may be considered as a social phenomena led by its current political, economic, cultural, and social environments. Therefore innovative frameworks, both theoretical and methodological, are expected to be designed specifically for studying social phenomena in contemporary China.

Furthermore, social science in China is experiencing rapid changes along with the society. Nowadays empirical-based research is developing and gaining importance among social scientists over traditional opinion-based, ideological and speculative duscussions. This trend is inevitable. Either the public, government, or academy welcomes higher quality empirical studies, which would serve their interests. What is going on in China goes beyond the pulic, as well as social scientists' understanding, therefore people are interested in learning changes that impact mostly on their daily lives. Besides, Chinese policy makers would benefit from having more accurate information and evidence in order to make more rational policy decisions. Finally, China is one of the world's most powerful countries with its important contributions to arts, sports, finance, natural science, technology, and world peace. We sincerely hope that our empirical social studies in China could make significant contributions to social science, and receive world-wide recognition and appreciation

1.2.3. Design Goals of CFPS

To conclude, the massive social transformation in China raised both challenges and opportunities for today's social scientists. To completely understand the society, we must understand such social changes. History calls for empirical social science research, and empirical research relies on high quality survey data. From a long-term perspective, it is impossible for contemporary social scientists to fully understand the ongoing social changes in China. That is exactly why we are making every effort to collect empirical materials to help future social scientists to better understand today's China. Thus, for both current empirical research in social sciences and future research

into present matters, we should value this unique opportunity to document the historical changes in China. Such is the central mission of the CFPS. With this in mind, we design the CFPS based upon the following three social characteristics:

First, social phenomena have multiple domains. Different domains are closely related rather than separated. Many aspects of social life, such as family background, social network, housing, education, occupation, income, and health, are interdependent.

Second, social phenomena have multiple levels. At the macro level, there are state policies, cultural traditions, and historical events; at the intermediate level, there are regional economy, urban facilities, and community environment; at the micro level, there are family structure, intergenerational relations, individual education and career, etc. All these factors at different levels contribute to individuals' life course.

Finally, social phenomena are continuous in time. Present behaviors depend on past events, and future decisions depend on today's experience. In fact, those seemingly simple social matters, such as population migration and family expenditures, are consequences of complex causal connections and dynamic processes.

Based on the multi-dimensional and multi-level nature of social phenomena, CFPS investigates at three levels: at the community level, we collect macro and intermediate data on political environment, rural landscape, infrastructure, population, resources, transportation, health system, and fiscal conditions; at the family level, we collect various information on family structure and relationships, living conditions, social networks, income and expenditure and family assets; at the individual level, we collect personal information on education, occupation, income, physical and mental conditions and marriage. In this way, the subjects in our study are connected and no longer isolated. Researchers may study their relationships between individuals, families and the society.

The continuity of social phenomena makes the social system more complicated and more difficult to study. Time is an important concept in studies of social issues and changes. From the perspective of methodology, time is a kind of information to identify the chronological order of events. For instance, individual behavior varies over personal experiences, which are closely related to time.[6] A panel study tracks these dynamics and captures their variation on the temporal dimension, which makes it a highly effective approach to study social trends. By observing the same group (the same sample) at different time points, we are able to study different individuals in different time spans in order to better understand causal processes over time and predict population trend. Thus, panel survey is of great value in such important areas of research as population heterogeneity, causal inference, and status transitions in social sciences.[7] Despite its high cost, complex designs and difficult implementation,

[6] Xie (2006)
[7] Ren & Xie (2011)

the panel survey provides more information and more valuable materials than cross-sectional and trend studies, thus brings more rewards for scientific studies. Because of this, CFPS decided to collect data of the target group at different time points in the first place. That is to say, we will keep track of the gene members as well as their biological or adopted children in CFPS 2010 baseline survey.

The data of CFPS is collected in a multi-dimensional, multi-level and longitudinal way, aiming to provide the most thorough and reliable data for more valuable and scientific research as well as to provide reliable evidence for state policy making.

1.2.4 Research Unit

The nature of social science research is heterogeneity.[8] Individuals, even in the same pool, differ from each other in characteristics. It is commonly known that education level, income, life style, physical condition, and social network vary among the population in China. The existence of heterogeneity leads us to distinguish between individuals. An individual is a basic social unit of human society heterogeneity, a source of social phenomena, such as health, happiness, and work. Thus in order to understand the society, we need to understand the differences in quality of life, financial status, and social roles between individuals. That is why an individual is our basic, and the most important study unit in CFPS, as well as the long-term follow-up survey target.

Family genes and background are not the only reasons that cause the heterogeneity in quality of life, financial status, and social role among the population. In addition, social environment and personal experiences play an important part. Thus, individual differences are affected by social structure, the most important one being the family.

First of all, the family is the principal environment for the socialization of the children. A person's ascribed status is determined by family, and family plays a major role in children's enculturation with social rules since born. The family influence on personal attitude, action, and expectation lasts in the later course. Thus, it is essential to understand a person's family in order to study the the person's later outcomes.

Second, family is a primary site for economic and social interactionsin Chinese culture. Many important social interactions, such as financial activity, housing, raising children, and supporting for the elderly, occur at family level. Thus, research on family-related topics is essential for studying Chinese society.

In addition, family affects relations between generations. How does parents' social status affect the next generation? How are family resources allocated between

[8] Xie (2012)

children? How are resources from grown-up children transferred to parents? A clear and comprehensive knowledge about the family and its members is the key to answering these questions.

Furthermore, family is crucial for research on marriage and gender. Marriage is common among grown-ups. Couples come from different families form new families through marriage or cohabitation, where social status and resources are redistributed and recomposed. In addition, marriage and family reflects gender differences in social capital and division of labor.

Last but not least, family is central to Chinese culture. Chinese people worshiped their ancestor and observed filial piety. They stove for achievements to glorify their family name. There was a strong norm for forming marriage between families of comparable social status. The desire for continuing the family lineage, along the male line, was paramount. Those traditional values indicate the importance of family in the Chinese culture. Family provides significant physical and emotional support for individuals. In return, people are obligated and responsible for paying the family back. Despite the waning of traditional family values in recent decades, the extraordinary significance of family in the Chinese culture remains. We can see this importance in parents' heavy investments in their children, extensive kinship networks, and reliance on informal transfers of financial resources among family members.

In summary, family is essential to understanding Chinese society, and thus an important research and investigation unit for CFPS. CFPS conducted an in-depth investigation of information on family relations and family members, and built an accurate family network structure which clearly established relatiohships between family members. Detailed information of family financial and social activities was also collected in our research. We do hope CFPS could provide a broader view and enhance understanding of Chinese society.

1.3. International Comparison

In the early stages of design, CFPS learned from some advanced research programs for their approaches and investigation tools, including PSID, NLSY, HRS, etc. In the meanwhile, CFPS has its own features and advantages which can meet various demands of researchers from different academic backgrounds.

Panel Study of Income Dynamics (PSID)[9] was designed and innitiated by the University of Michigan in 1968. It is the most authoritative panel study of family economy in the United States. The original purpose was to study poverty and the effect of Lyndon Johnson's "War on Poverty" project on economics and the public welfare. Later on, the research topics gradually expanded to employment, income,

[9] For more information on PSID, please refer to http://psidonline.isr.umich.edu/default.aspx.

wealth, housing, food expenditure, transfer payment, marriage and fertility. At the beginning, the sample size was 5,000 households. One of the adults in each household was chosen to be interviewed by telephone. Since 1997, the Child Development Supplement to the PSID (PSID-CDS) has been added to the former study.

National Longitudinal Surveys of Youth (NLSY)[10], designed by the Ohio State University and launched by the National Opinion Research Center (NORC) of the University of Chicago, is an authoritative panel survey program studying the American youth on the labor market. The longitudinal surveys involve two cohorts of young people. The 1979 cohort contains 12,868 respondents aged between 14 and 22 when they were first interviewd in 1979, and the 1997 cohort contains 9,000 young people aged between 12 and 16 when they were first interviewed in 1997. The research focuses on human capital of the youth and their labor market activities. The survey includes education, employment, career trainings, working hours, income and assets, attitude and behavior, health, and political involvement, etc.

Health and Retirement Study (HRS) started in 1992. Designed and launched by the University of Michigan, it was the most influential longitudinal program on aging in the United States. The survey targeted at old people aged above 50. The sample size is about 26,000. HRS concentrates on old people's labor force participation and health changes, as well as information about income, work, assets, pension plans, health insurance, disability, physical health and functioning, cognitive functioning, and health care expenditures.[11]

In general, PSID, NLSY and HRS are large nationally representative panel research programs in the United States. Although the mentioned studies are monographic studies, these research is comprehensive and offers data to researchers across different fields with different interests, which is an important data source for a variety of related scientific research. CFPS benefits from their experiences, and we hope to contribute to social research in China as well. Similar to the surveys above, CFPS consists of a large, nationally representative sample: 14,960 households were interviewed in the baseline survey, including 57,155 gene menbers for long-term follow up survey. As a multi-level, multi-dimensional, and longitudinal research program, CFPS covers a thorough and comprehensive content, including various important life events through lifetimes, as well as some specific designs on family relationships, family financials and communities.

Based on the experiences of the surveys above, CFPS has made some improvements. We believe the multi-level social structure is a critical characteristic of Chinese society, in which family is of crucial importance. Therefore, CFPS conducted a more extensive and intensive survey on family members and their family relationships. Surveys on family relationships usually focus on one or two core

[10] For more information on NLSY, please refer to http://www.bls.gov/nls/nlsy79.htm, and http://www.bls.gov/nls/nlsy97.htm.
[11] For more information on HRS, please refer to http://hrsonline.isr.umich.edu.

members of a household to collect information of their family background and other members in their family. The data collected in this way is rather limited. In contrast, CFPS asks all family members who are eligible for the research, even children, to complete their personal questionnaires either by themselves or with the help of others. Thus, the statistics provided by CFPS are more detailed, accurate and complete. Moreover, the unique design of CFPS[12] enables researchers to acquire information not only on the one-dimensional relationship between interviewees and other family members but also with the entire family; not only on immediate relationships but also cross-generational and sibling relationships; not only on members living in the household but also members living apart. With the help of CFPS individual coding system, researchers can locate family members and obtain an accurate family network. Thus, it is an important advantage of CFPS that it provides more valuable information on family structure and family members.

1.4. Survey Technologies

As a national comprehensive survey program, CFPS consists of a large sample size, a wide survey coverage, and a complicated design, which make traditional paper and pencil survey no longer feasible. Instead, a computer assisted interviewing technology was adopted to guarantee the efficiency and quality of the survey.

Computer assisted interviewing technology includes Computer Assisted Personal Interviewing (CAPI), Computer Assisted Telephone Interview (CATI), and Computer Assisted Website Interview (CAWI). CAPI was used in the CFPS baseline survey in 2010. Computer assisted interviews rely on the interview management system, which is a professional interviewing software. The system could assist interviewers in completing the questionnaire and managing interview data information. Moreover, it is more convenient for interviewers to contact the headquarters and receive feedbacks on problems. The main functions of the interview management system are as follows:

(1) Complicated designs of electronic questionnaires. Various questions using multiple choice, single selection, tables, loops, intervals can be applied. In the meantime, customized questions can be designed via complicated logical skips in accordance with the characteristics of different groups. With the setting of hard checks and soft checks, CAPI questionnaires can give prompt instructions for some illogic or unreasonable answers, which makes it possible for interviewers to communicate with respondents in time and make corrections.

(2) Quick management of samples. The headquarters can send samples to interviewers remotely and modify the interviewing tasks according to the practical needs of field survey. The system can also record detailed information on each interviewee and respondents family, including the address, interviewing method and payment, so as to provide convenience for supervisors and interviewers.

[12] It refers to the T table design. It is introduced in detail in the following part of the handbook.

(3) Real-time data transmission. With the help of the interviewing management system, interviewers can transmit and exchange data with the headquarters in time. The headquarters can easily track the progress and exercise remote control by spotting the problems in time and trying to solve them with interviewers at field. In addition, the interviewing management system saves the efforts of inputting data in traditional surveys using paper and pen; the data cleaning and analysis can be done on a contemporaneous basis.

(4) Real-time supervision of interviewer behavior and quality control. The system can supervise the computer operations of interviewers via recording and the headquarters can inform the interviewers to improve their non-standard interviews in time.

(5) Paradata analysis. The interviewing management system is able to collect a set of paradata, such as pause time for every question, records of interviewers' modifications of the options, etc. The analysis of these data will provide scientific evidence for the improvement of survey research design in the future.

2. Sampling

2.1 Sampling Design

The sample of CFPS is drawn from 25 provinces/cities/autonomous regions in China excluding Hong Kong, Macao, Taiwan, Xinjiang, Qinghai, Inner Mongolia, Ningxia, and Hainan. The population of these 25 provinces/cities/autonomous regions in China (excluding Hong Kong, Macao, and Taiwan) covers 95% of the Chinese total population. Thus, CPFS can be regarded as a nationally representative sample.

The original target sample size was 16,000 households. Half of the sample (8,000) was generated by oversampling with five independent sampling frames (called "large provinces") of Shanghai, Liaoning, Henan, Gansu, and Guangdong. Each of the subsample had 1,600 households. The other half 8,000 housholds were from an independent sampling frame composed of 20 provinces (called "small provinces"). The "large provinces" were representative at the regional level, which could contribute to provincial population inferences and cross-region comparisons. With a second-stage sampling, the five "large provinces", together with "small provinces", made up the overall sampling frame representative of the national population.[13]

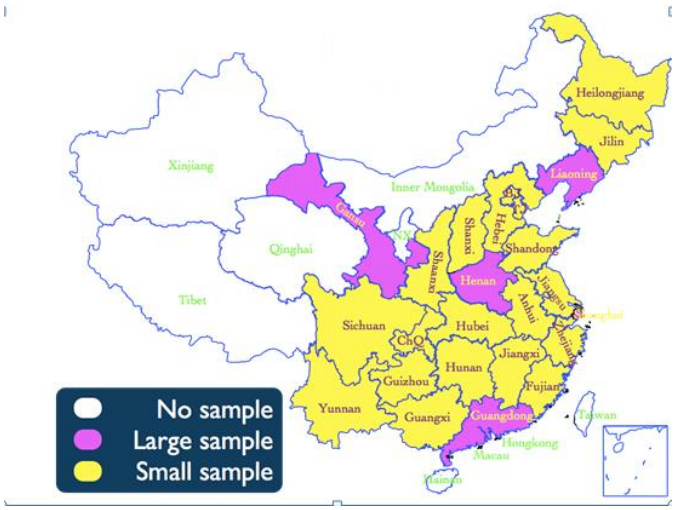


Figure 4. The Sources of CFPS Samples at the Provincial Level

[13] We call this sample "national re-sample" or "constructed sample".

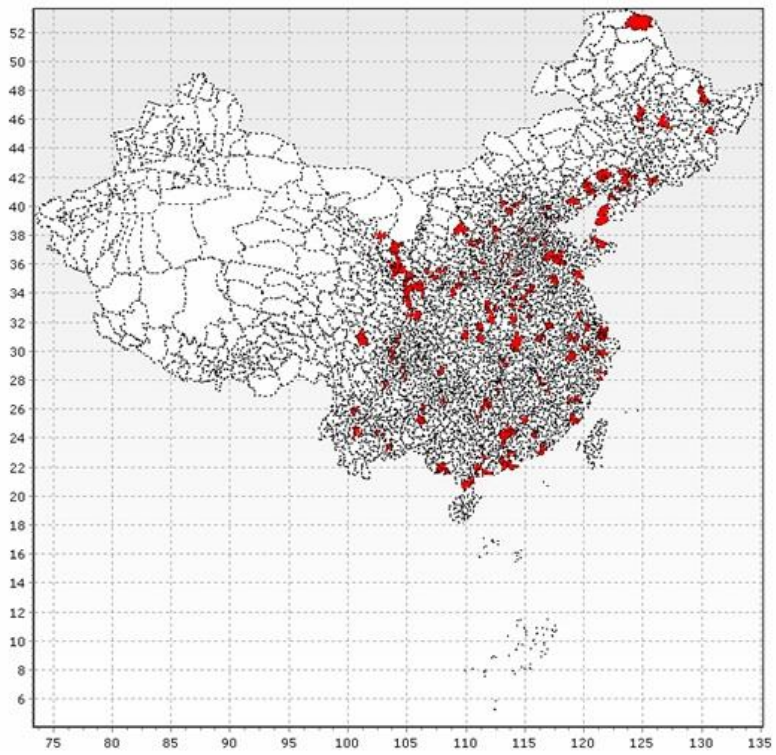


Figure 5. The Geographic Distribution of CFPS Samples at the County Level

Table 1. CFPS Target Sample Size

Category	Provinces/Cities/Autonomous Regions	Target Households	Oversampling Rate
"Large Provinces"	Shanghai	1600	10.28
	Liaoning	1600	4.45
	Henan	1600	2.04
	Gansu	1600	7.30
	Guangdong	1600	2.02
"Small Provinces"	Jiangsu, Zhejiang, Fujian, Jiangxi, Anhui, Shandong, Hebei, Shanxi, Jilin, Heilongjiang, Guangxi, Hubei, Hunan, Sichuan, Guizhou, Yunnan, Tianjin, Beijing, Chongqing, Shaanxi.	8000	1.00

Taking the reginal differences in Chinese society and the reduction of survey processing cost into consideration, CFPS implemented Probability-Proportional-to-Size Sampling (PPS) with implicit stratification. Administrative units and socioeconomic status (SES) are used as the main stratification variables. Within the administrative unit, local GDP per capita was used as the ordering index for SES. If the GDP per capita in the administrative unit is not available, the proportion of non-agricultural population or population density is used alternatively.

All the sub-sampling frames of CFPS were obtained through three stages: the Primary Sampling Unit (PSU) was administrative districts/counties, the Second-stage

Sampling Unit (SSU) was administrative villages/neighborhood communities, and the third-stage (Ultimate) Sampling Unit (TSU) was households.[14] In the first and Second stages, CFPS used official administrative divisions for the sample selection. The third sampling stage is a systematic selection of housing units from street listing with random starting point and equal probability method. For the consideration of response rate, the 2010 CFPS survey used the estimated response rates from 2008 and 2009 pilot studies as reference and enlarged the sample size proportionately. A total of 19,986 households were selected according to systematic sampling principles, which ensuredthe expected sample size for the survey.It is important to note that CFPS sampled the Chinese population as a whole instead of using traditional sampling design, which sampled urban and rural areas seperately. The reason behind is that the official rural-urban division can hardly reflect the reality of China's rapid urbanization. At the community level, we collected information regarding whether the sampled communities were urban neighborhood or village. At the household and individual level, we identified the *hukou* type and whether the household engaged in non-agricultural work or not. Users may decide whether it is a rural or urban area using such information rather than administrative divisions.

For detailed information of sampling design and implementation, see *Sampling Design (CFPS-1).*

Table 2. Three Stages of CFPS Sampling

Stages	Guangdong, Gansu, Liaoning, Henan: 4 "Large Provinces"	Shanghai: "Large Province"	"Small Provinces"	Total
Primary	4×16 Counties=64 Counties	32 Streets (Towns)	80 Counties	144 Sampled Counties+32 Sampled Streets (Towns)
Second-stage	64×4 Communities =256 Communities	32×2 Communities =64 Communities	80×4 Communities =320 Communities	640 Communities
Third-stage	640×[28, 42] Households			19986 Households

[14] Shanghai is different from other "large provinces", and therefore the sampling procedure was slightly different. More information can be found the technical report: CFPS-1.

Table 3. Terminal Sample Size of CFPS 2010 Baseline Survey[15]

Region	Category	Expected Response Rate	Contacted Sample Size
Low Response Rate	neighborhood committee (urban area and village in urban fringe[16])	60%	42
	Other village	70%	36
Intermediate Response Rate	Neighborhood committee (urban area and village in urban fringe)	70%	36
	Other village	80%	32
High Response Rate	Neighborhood committee	80%	32
	Village	90%	28

2.2 Terminal Sampling Frame

Given high population mobility and separation of residents from registered household in China, we believe that the sampling based on the rosters of urban neighborhoods and villages would cause a large number of omissions. In order to get a complete sampling frame covering all the residents and improve the accuracy of third-stage sampling, we drew the map of local villages with paper and pen in the field. Several pilot testings were done in four villages in Beijing and Hebei from the beginning of 2009 to Agust, 2009. With detailed information on the features and coding system of buildings, the method of household list making, current map and the usability of households list, we designed the preliminary plan for village sampling frame. We did pre-tests in 4 urban neighborhoods/villages in Gansu and 4 in Zhejiang in November and December of 2009 to examine the practicability of the design. After that, we summarized the experience and improved the mapping and street listing design.

From December, 2009 to April, 2010, we conducted 23 group training for mapping and street listing. Each group was trained for 3 days and there were 243 draftsmen trained in total. The training included the drawing and coding of the buildings, collecting auxiliary materials, making building list, and household list, etc.

The mapping officially started in December, 2009 and finished in June, 2010, with the maps, statistical tables of residents' information, and households' list of 649 urban neighborhood/villages done. Different methods were applied to aduit the maps and other materials in order to ensure quality and credibility. For detailed map making methods and verification standards, see *Third-stage Sampling Frame Construction (CFPS-2).*

[15] Technical report: CFPS-1.
[16] The division of main districts and urban fringe refers to the urban/rural code by the Department of Statistical Design & Management, National Bureau of Statistics of China.

After organizing these map materials and coping with some special problems such as multiple residences for one household, multiple households for one residence and failure to confirm the address,[17] the group started the third-stage sampling.

[17] See Technical Report CFPS-1 for more details.

3. Questionnaire Design[18]

3.1 Overview

In 2010, CFPS did surveys at three levels. At the community level, CFPS did an overall interview of the sampled villages/urban communitiesmainly focused on the infrastructure, population structure, policy implementation, economy, and social service, etc. At the family level, after eliminating invalid addresses, non-residential structures, and vacant houses, CFPS selected one member of each eligible household to fill out two questionnaires, one on the family members' basic information and members' relationships, and the other on the basic information of the whole family. At the individual level, for eligible family members, children under 16 answered child questionnaires and those older than 16 answered adult questionnaires.

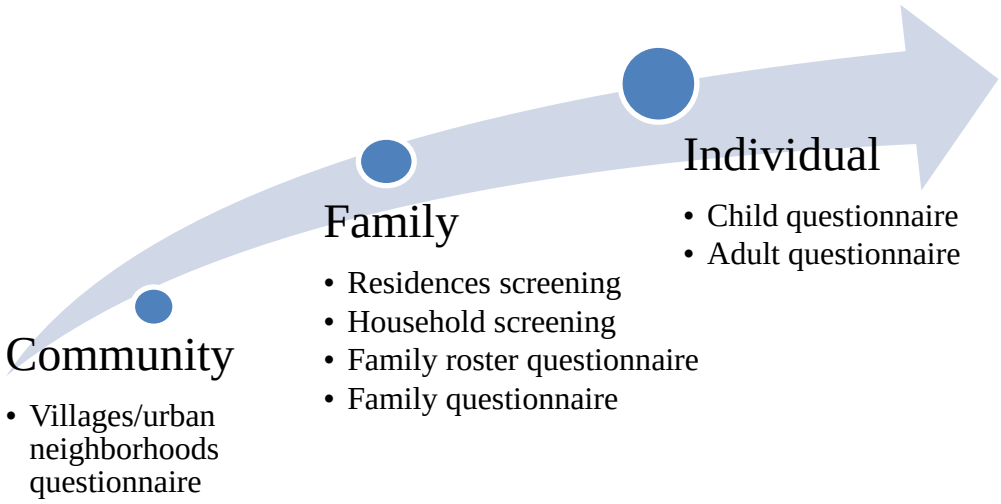


Figure 6. The Levels of the Questionnaires of CFPS 2010

Five major questionniares were designed in 2010 CFPS: the community questionnaire, the family roster questionnaire, the family questionnaire, the child questionnaire and the adult questionnaire. The child questionnaire was divided into two types: for children under 10, their guardians would answer partial questionnairefor them; for children aged between 10 and 15, apart from the guardians' part, they need to answer some questions by themselves.

Besides the five major quesionnaires, CFPS also designed residence screening and household member screening quesinnaires. The household screening

[18] For more detailed information on questionnaire design, please refer to Sun et al. (2011).

questionnaires were collected through the interviewers' observations and judgments in the field surveys; schematized questionnaires were not needed. The household member screening, similar to tranditional questioannires, was finished by interviewing the household members.

CFPS applied multi-modules design for questionnaires. Each questionnaire was composed of different modules according to different situations of families and individuals interviewed. The CAPI system provided much convenience in setting up personalized questionnaires in interviewing. For example, we clicked the school module for those still at school and clicked work module for those employed. It is also the reason why we did not use two different questionnaires specifically in rural and urban areas.

Figure 7 is the flow chart of all the questionnaires of the baseline survey 2010.

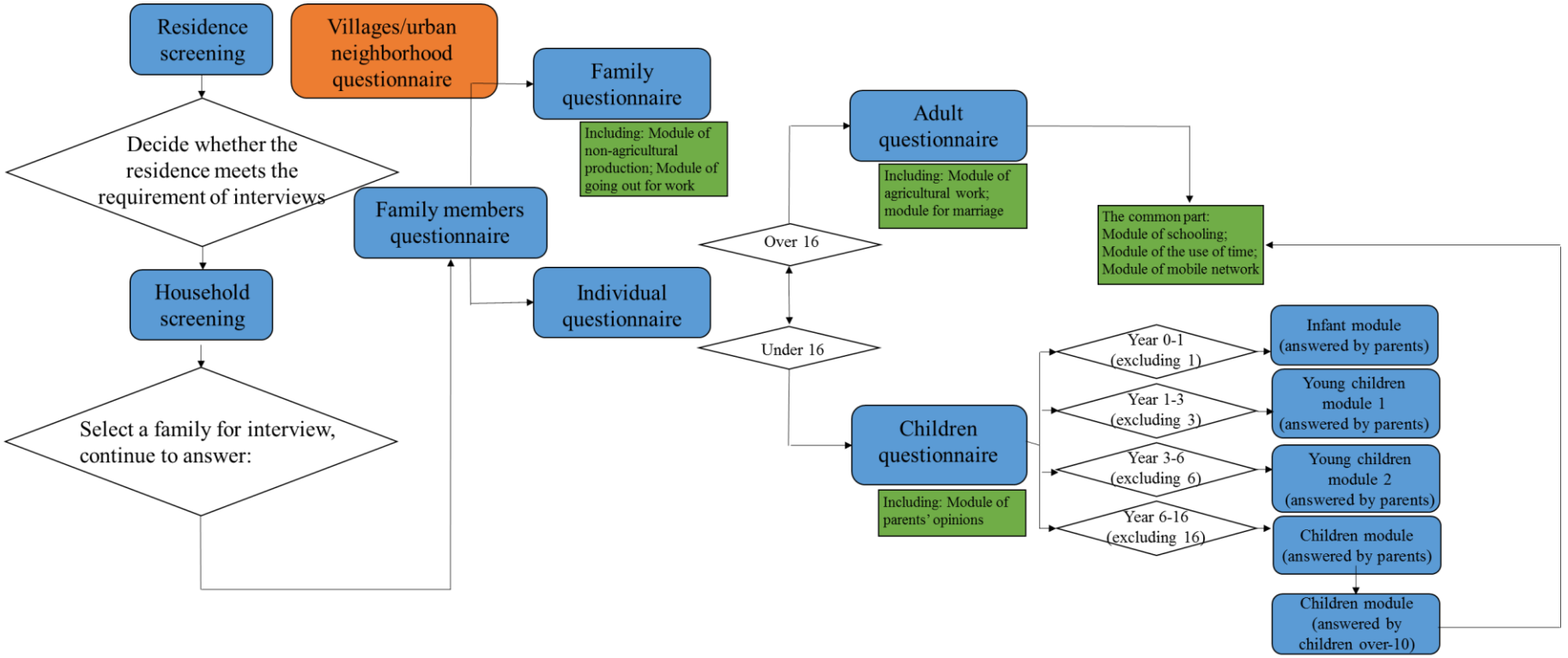


Figure 7. The Structure of Questionnaires of CFPS Baseline Survey 2010

3.2 Community questionnaire

The goal of the villages/urban community questionnaire is to acquire the information on the infrastructure, population, politics, economy, history, policies, and related information of the villages (i.e., rural communities) and urban neighborhoods (i.e., urban communities). For the process and content, see Figure 8 and Table 4.

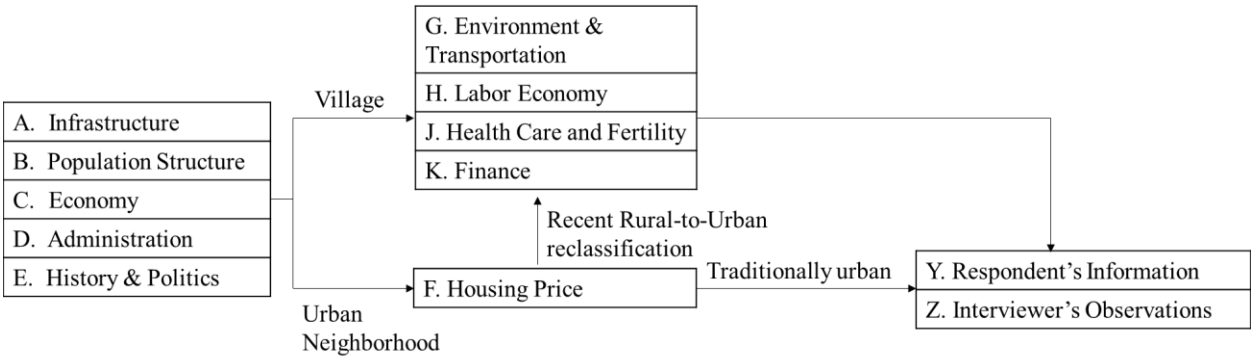


Figure 8. The Flow Chart of the Community Questionnaire of CFPS 2010

Table 4. Main Content of Community questionnaire of CFPS 2010

Module	Questionnaire Content
A. Infrastructure	type of village/urban neighborhood, respondent's position, facilities, bulletin, land borders, administrative area, water sources, fuels
B. Population Structure	Total number of households, population, registered residents, permanent residents, migrant population, age structure, fertility and mortality, ethnic groups
C. Economy	basic living allowances, price level
D. Administration	type of administrative staff, working conditions, neighborhood transportation
E. History & Politics	historical changes, whether it is a tourist area, whether there are high-pollution enterprises, latest election of village/urban neighborhood's' committee
F. Housing Price	highest price in history, highest last month, average price last month
G. Environment & Transportation	transportation hours to the closest county, town, provincial capital, mineral resources, natural disasters, land resources
H. Labor Economy	labor force structure, gross output of agriculture & non-agriculture sectors, net personal income per capita, price of assisting workers, distribution of the "big family names"
J. Health Care & Fertility	area of medical station, number of medical staff, progress of rural cooperative medical services, one-child policy
K. Finance	Collectively-owned enterprises and their output, total fiscal revenue and its resources, total financial expenditure and the items
Y. Respondent's information	gender, age, political status, level of education of the respondent, information of the director of the village/urban neighborhood, names of

	professions of other respondents
Z. Interviewer' Observations	economic status, tidiness of roads, mental state of villagers/neighborhood, the homogeneity of villagers/neighborhood, architecture pattern, congestion, type of village/ neighborhood, topography, features of respondents

For the interviews using the village/urban community questionnaire, we ask interviewers to reach those who know the communities well and have access to statistical data. The staff members of the village/ neighborhood committee, especially the director who is in charge of daily administration, would be the best information source. Other staffs such as the accountants in the village/ neighborhoods who also know well about the community because of their job or their long years of service can also be potential respondents. Moreover, the secretaries of the Party branches in the village/neighborhood can also be our respondents if they have a comprehensive knowledge of the village/neighborhood. When the first type of respondents cannot answer all the questions, it is allowed to ask the second and the third type of respondents to cooperate and fill out the questionnaire. But these are individual interviews in separate rooms rather than collective interviews.

While using the family roster questionnaire, "the end of the year" refers to the last day of a natural year. For example, the end of the year 2009 refers to December 31st, 2009.

3.3. Residence Screening

In order to ensure the accuracy of the terminal sampling, although our draftsmen have already eliminated the vacant houses and the non-family households by field interviewing, consulting the neighbors and staff members of the village/neighborhood's committee during the construction of the ultimate sampling frame, in actual interviews, it is still required to find the corresponding samples selected according to the map addresses and confirm that they actually are residences before starting interviews. The confirmation of houses and residences is called the residence screening.

The interviewers shall check whether the sampled addresses according to our map are valid—that is, whether the addresses provided actually exist. After verifying the information, they shall decide the types of the building on the addresses by consulting the residents or others in the know. Qulified residentce sample shall go through the screening. Invalid addresses, non-residences, and vacant houses shall be eliminated from our interviews. For the samples that are difficult to confirm, multiple attempts are required to gain the accurate information.

3.4 Household Screening Questionnaire

After confirming the residences on the sampled addresses, the survey goes through the household screening process. The main purpose of the questionnaire for the screening of residents is to pick out the households that meet with the requirements of interviews within certain residences. Below are the procedures of the screening questionnaire:

First, determine the number of independent economic units within the sampled residence. For example, for parents and children living together, if they compose one economic community, we define them as one independent economic unit; if they are two economic communities, we define them as two units. The defination of the economic units is not limited to the house owners, but also includes those who have at least a part of the right of habitation of the sampled residences.

Second, determine the eligible family households among the independent economic units. Two conditions below are not qualified for survey:[19]

1. For those independent economic units that are only composed of one person, who at the same time belongs to an economic community with more than two other family members elsewhere, we do not regard them as "households" in our survey. Instead, he or she is seen as one of the family members of that family household he or she belongs elsewhere.

2. We require all the households belong to China mainland, which means there must be at least one family member of Chinese nationality (excluding Hong Kong, Macao and Taiwan).

Finally, if there are more than one household on one specific residence address that meet the requirements above, the computer system shall automatically run the program of random sampling, which will select one of the households to be interviewed. For all the households on the address, the questionnaire for the screening of households includes questions on the houses they own outside the local village/neighborhood and elsewhere in China.

3.5 Family roster questionnaire

3.5.1 Design Purpose

The family roster questionnaire is designed to collect information on the family structure, the relation network of immediate relatives, and all the basic information of family members and their immediate relatives, including gender, age, educational level, occupation, administrative posts, hukou, and residence, etc. Meanwhile, the family roster questionnaire lays a foundation for individual questionnaires: we can acquire information on the roles of different groups in our survey from the family roster questionnaire so as to decide which type of individual questionnaires they should complete. Moreover, the definition of family members is the basis of our follow-up surveys.

3.5.2 The identification of family members

The identification of family members is shown in the following flow chart (Figure 9).

[19] The original design of the screening of households required at least one of the family members who had lived in the sampled residence for 6 months. But this was abandoned executing process, for it could only eliminate very few households.

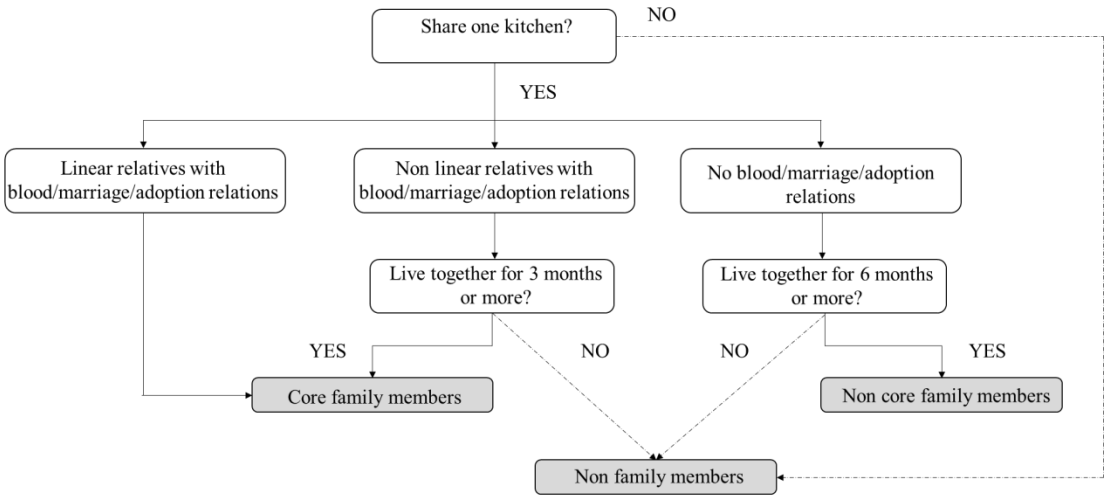


Figure 9. TheFlow Chart of the Identification of Family Members[20]

In general, conditional on "sharing one kitchen", the family members of CFPS are divided into three types. The first type is all the immediate relatives, the second type is non-immediate relatives who have been living in the family continuously for 3 months or longer up till the interview, and the third type is those who have no blood/marital/adoptive relationship with the family but have been living in the family continuously for 6 months or longer up till the interview. The first and second types are the core family members. They together with their biological and adopted children are the gene members of CFPS, which are the permanent respondents of CFPS until their deaths. The third type includes non-core family members. Once they leave the sampled families, we no longer include them in our follow-up survey.

3.5.3 Content of the Family Member Questionnaire

The respondents of the family roster questionnaire must meet two requirements: they are among the members who are sharing one kitchen and they have blood/marital/adoptive relationships with the family. The specific interviews of the questionnaires are done in the following way:

[20] This figure has been modified on the basis of Sun et al. (2011, p.126).

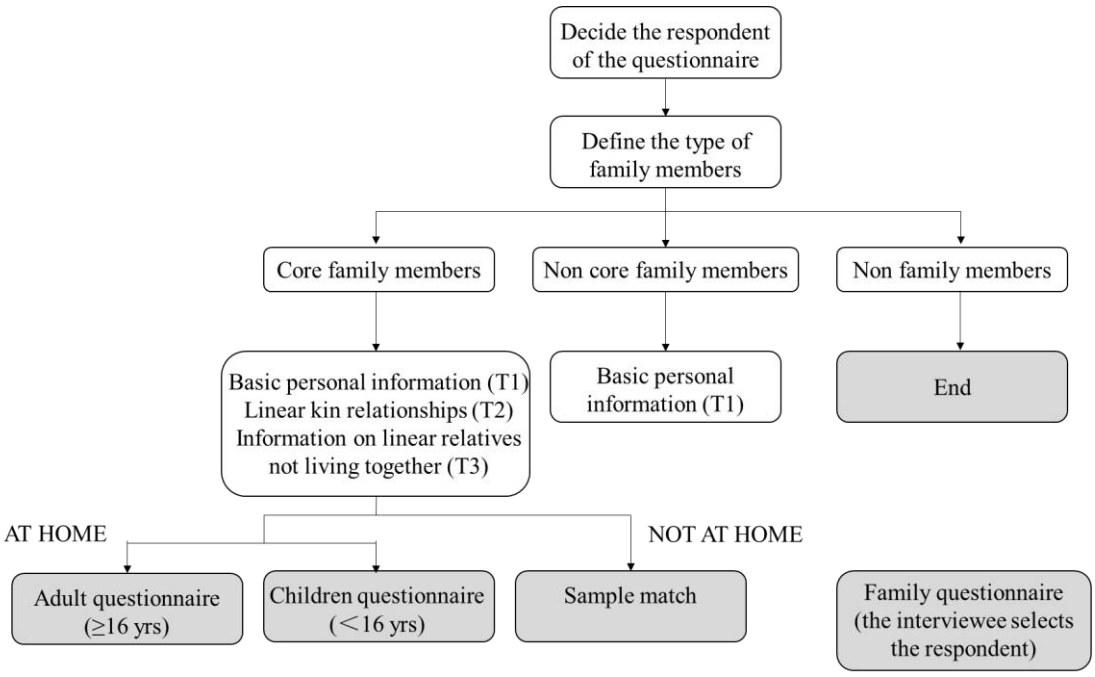


Figure 10. Flow Chart of Interviews Using the Family Member Questionnaire[21]

For each core family member, the questionnaire should include basic information of not only him/herself, but also of his/her direct relatives. For non-core family members, we do not collect the basic information of their immediate relativeimmediate relatives. When the family roster questionnaire is completed, the system will select the specific individual questionnaires for every core family member present according to their age; for those core family members who are not present, they automatically go to the matching system. The non-core family members defined by the family roster questionnaire do not complete further individual questionnaires.

According to the design of CFPS, the interviews of the core family members who are not living at home are still required, which means that they need to complete the child/adult questionnaire accordingly. There are four types of exceptions however: those who have become a monk or a nun, those who are in prison, those who are servicing in the army, and those who have moved abroad. If the core family members are not presented because of these four reasons above, the interviews are not required to compelte the interview in that year.

In the interviews in 2010, we have only done the local research, namely the interviews of the sampled households and family members within the village/neighborhood and the individuals who were still in the local districts/counties. We did not interview the core family members who were out of the sampled districts/counties. Their basic information was collected through our interviews with their relatives, friends, and neighbors.

[21] This figure has been modified on the basis of Sun et al. (2011, p.126).

3.5.4 The Design and Implementation of the T Tables

The design of the T tables of the CFPS family roster questionnaire is an original method to collect information on family relations.

As we know, in past social surveys, the information on family relations was collected through questionnaires at the individual level, which directly ask the respondent to provide information about their parents, spouses, and children. There are several drawbacks of such approach:

First, these surveys normally select only one respondent as the center of family relations and ask questions on his or her specific relative relations. It is thus presumed that there is only one center of family relations, i.e., the respondent himself/herself. Normally, a random assignement is given within the household selection. An alternative way is to select the household head. However, as every family member can be treated as the family center, the selection of center might not be unique yet meaningful. The family relationship should be more like a family tree network with multiple centers. Previous methods only build up a radiation structure from a single core (e.g., the household owner or the respondent), which is only a small part of the familytree network.

Second, since the traditional methods only collect radiation structions based on single cores treating the respondents as the centers, researchers can only learn about the relations between the respondents and other family members/relatives whereas the relationships among family memebers apart from the repondents remain unknown.

Third, previous surveys only ask questions about the "father", "mother", and "children" of the respondent, without asking their names or coding these relatives. Thus, even if more than one family member is interviewed, it is impossible to relate them in one family network through their names or codes.

Fourth, the information collected from previous surveys is mainly of the same generation (e.g. siblings and spouses) or successive two generations (e.g., parents and children). Without the information on the relations between relatives and their names, the information on the relations across generations is unknown.[22]

The T tables of CFPS solve the above problems. The T tables are composed of three tables T1, T2, and T3, lying in the beginning part of the family roster questionnaire (see Figure 11). Table T1 (the table on family members living together) and table T3 (the table on immediate relativeimmediate relatives not living together) record the basic social and demographic features of every family member and their immediate relatives (referring to their parents, children and spouses) who are not living with them. Table T2 (the table on relations) identifies the relations of all the family members and the corresponding relations between T1 and T3 members.

[22] See Technical Report: CFPS-7.

The Table on Family Mmembers Living Together (T1)

Personal code	Name	DOB	Sex	Marital status	Highest degree	Main job	Admin./Managerial position	Info.on people outside home
101								
102								
…								
301								
302								
…								

The Table on the Relations of Immediate Relatives (T2)

Personal code	Name	Father	Mother	Spouse	Child 1	Child 2	…	Child 9	Child 10
101									
102									
…									

The Table on Linear Family Members Not Living Together (T3)

Personal Code	Name	DOB	Sex	Alive or not	Marital status	Highest degree	Main job	Admin./ Managerial position	Info. on residence and *hukou*
301									
302									
…									

Figure 11. The Design of the T Tables

All the information of the T tables is completed by the respondents of the family roster questionnaire. It does not require every family member to get involved in filling out the tables. In the field interviews, the interviewers complete the T tables with the assistance of CAPI.

First, in the introduction to the family roster questionnaire above, we have already known that the family members living together are identified by a set of questions on the basis of economic relations. Those who are qualified as family members living together in this process shall fill in the table T1 sequentially. The final T1 table includes the core family members with individual codes starting with the number “1” and the non-core family members with codes starting with “3”.

Next, according to the information of the core family members in table T1, the original name list of table T2 is automatically generated. The table T2 is completed in a way of “taking-turns”, which means that everyone takes turns to confirm their relationships with other immediate relatives (parents, children and spouses) as the center of family relations. That’s how the table on the relations of immediate relatives (table T2) is done.

Finally, we collect the basic information of all the immediate relatives of the family members included in the name list of the table T2. For the immediate relatives mentioned in T2, if they already exist in the table T1 (i.e., they are the family members living together), the CAPI system will automatically add their information; there will be no repetition of interviews. If the family members do not exist in T1, the system will generate a name list for the T3 table. Again, the interviews will be done sequentailly until the final table on immediate family members living together (table T3) is done. The individual codes of all the members of T3 will start with the number "2".

As such these three tables jointly present a thorough family network, where the relations within the same generation, immediate generations, and gapped generations are all connected, with detailed records of each member's personal information. The design of the T tables fixes the disadvantages of previous surveys in the inaccurate information on family relations, providing richer information on an overall picture of family relations. It also provides valid information of all family members for in-depth studies on family issues. Meanwhile, the specific implementation of the T table avoids repetition of interviews, which greatly increases the efficiency of the survey.

Despite the multiple advantages of the design of the T tables and maximal considerations of convenience and efficiency issues, the surveys cost more time and are more complicated compared with regular surveys due to the large amount of information collected. The behaviors and interview skills of the interviewers have a direct influence on the cooperation of the respondents and the quality of the T tables. Therefore, strict and comprehensive trainings are needed for the interviewers before the field surveys.

3.6 Family Questionnaire

The main purpose of the family questionnaire is to collect information on daily life, social interactions, and economy of the sampled families. The main content of the family questionnaire is shown in Table 5.

The family questionnaire is answered by the household head who has comprehensive knowledge of detailed family life and financial situation It is preferred that one family member completes all the questions. However, under special circumstances, it is also allowed to change the respondent, for example, when the respondent has to leave in the middle of the interviews, other family members can answer the rest of the questions instead of him/her. But it is required that the substitute also knows well about matters in the family.

Table 5. Main Content of the Family Questionnaire of CFPS 2010

Module	Content
A. Geography and transportation	Nearest public transportation, medical stations, high schools, commercial centers
B. Living conditions	Use of water, fuel resources, electricity, bathroom conditions, garbage disposal, employment of house maids
C. Social interactions	Spring Festival visits, gift giving, family-lineage, ancestor worship, neighborhood interactions, communications with relatives
D. Housing	House ownership, self-built houses or commercial apartments, houses for rent, the house area, living time, the market value and the rent of the houses, the apartment structure, other estates, the difficulty in housing
E. Management conditions[23]	U module for working outside (people who are outside, working address, time devoted, whether they go home during vacations, the transfer payment, whether their family has increased/decreased assisting workers because their working outside), the government's support, the reason of poverty, V module for non-agricultural management (the type of the non-agricultural industries, the number of participants, the total assets, the shares held by family members, the number of employees, the turnover, after-tax profits), houses for rent, land and other means of production for rent, property for sell, the demolition of the houses, the land acquisition
F. Family income	Savings, financial products, pensions/social security/basic living allowances, salaries/rewards/allowances/bonuses, non-salary/agricultural income, value of gifts
G Family assets	Insurance indemnity, others' debts, the value of collections, the present value of other assets
H. Family expenditure	The most expensive expenditure, loans, daily expenditure (food, travel, communications, etc.), special expenditure (family appliances, medical care, education, commercial insurances), donations, total expenditure
J. Durable goods	Cars, motorcycles, tractors, televisions
K. Agricultural production	The land type, land areas, revenue & expenditure, the type of crops, output, sales, income, the type of domestic animals and fishing, its output, sales, income, the raising conditions of domestic animals
Z. Interviewers Observations	The answerer's housing situation, the tidiness of the family, the mental state of their family members, the relations of family members , between the elder and the young and between genders, personal characteristics of the respondent

[23] The E module consists of U (working outside) and V (non-agricultural management). If there are family members going out for work or work in non-agricultural industries, they turn to answer sub-module U or V. After finishing the questions in these two parts, they come back to the E module and continue the following questions.

3.7 Individual Questionnaire

In CFPS, those younger than 16 are defined as children and those over 16 are defined as adults. The child questionnaire and the adult questionnaire are the two types of individual questionnaires designed for these two groups of people.

Before introducing the content of the individual questionnaires, we need to give an introduction to the principles of our design.

First is the calculation method of age. As the age of 16 is the dividing point for children and adults and the questionnaires for children have different modules for children of different ages, the calculation method of age is very important. The CAPI system automatically creates the individual questionnaires based on the date of birth or the age of the respondents. While calculating the age, the system only considers the year of birth, omitting the month. Specifically, the age equals the year of the survey minus the year of birth. For instance, for a child who was born in October 2000, although he was 3 months away from turning 10 in July 2010 when we were doing the survey, we regarded him as 10 years old in CFPS. This rule is also applied in the other parts of CFPS questionnaires and follow-up surveys.

Second is the respondent of the questionnaires. The adult questionnaire must be completed by the adults themselves. The child questionnaire is divided into two parts: the part completed by children themselves and the part completed with others' help. Among them, there are 4 types of others-answering questionnaires according to the age of the children: 0 (excluding 1 year old), 1-3 (excluding 3 years old), 3-6 (excluding 6 years old) and 6-16 (excluding 16 years old). Children between 10 and 16 (excluding 16 years old) need to finish another questionnaire by themselves apart from the part answered with the help of others. The additional questionnaire must be completed by children themselves independently. The part answered by others is required to be completed by their parents. Here "parents" refer to the main guardians living with the children interviewed, who take the most care of them and know the best about them. Moreover, we require the respondents to have the language and thinking abilities. If the sampled respondents have difficulties in language or thinking, e.g., mentally retarded, with serious mental diseases, or suffering from Alzheimer's, we apply the corresponding special codes and stop interviewing individuals. For these particular respondents, users can obtain their basic information from the table T1 in the family roster questionnaire.

The detailed structural process and the content of the interviews of the child questionnaire are shown in Figure 12. In the part answered with others' help, the number of questions included increases with the age of the children. It can be seen from the figure that children between 0 and 1 have the least amount of questions to answer. Additional information is to be collected for children between 6 and 16: apart from all the questions for children between 0 and 1, 1 and 3, 3 and 6, we still need to interview about their schooling and caring from their parents. We also need to ask their parents about their judgment of their children's behaviors and habits.

The adult questionnaire is relatively more straightforward than the questionnaires for children. There are less logical skips between modules, see Table 6 for detailed content. Some modules have relatively complicated inner structures and we list the most complicated one in the following part (see Figure 13) to make it easier for users to understand.

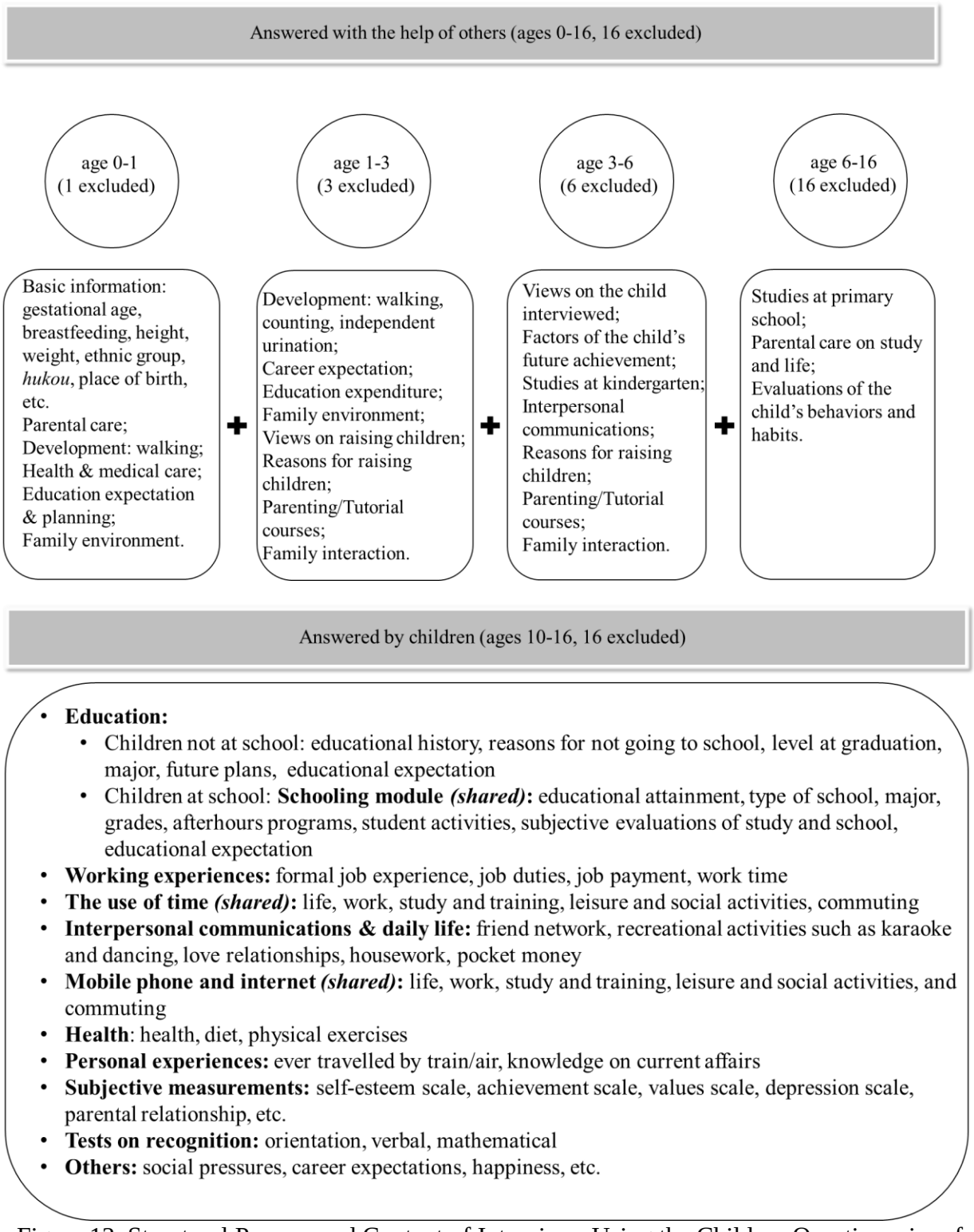


Figure 12. Structural Process and Content of Interviews Using the Children Questionnaire of CFPS 2010[24]

[24] This figure has been modified on the basis of Sun et al. (2011, p.151).

Table 6. Main Content of the Adult Questionnaire of CFPS 2010

Module	Content
Basic information	Date of birth, birth weight, birth place, residence, *hukou*, ethnicity, family category during the Cultural Revolution, political party and organizations affiliations, time spent living with parents before age 3 and between ages 4 and 12
Siblings	Number of siblings, name, date of birth, alive or not, age at and cause of death, marital status, educational attainment, occupation, administrative/managerial positions, residence, living with parents, parents' cause of death
Educational history	Educational attainment, type of schools at different educational levels, time spent at school, when did school end, name of schools, graduated or not, subject and major, educational expectation
Language use	Importance of different languages, language spoken at home
Schooling *(shared)*	Current level of education, type of the school, major, grades, extracurricular tutoring, student activities, subjective evaluations of study and school, educational expectation
Marriage	Current marital status (married/remarried/cohabitating/divorced/widowed), date of birth of the present/last/first spouse/cohabitation partner, time at marriage/cohabitation, pre-marital cohabitation, channel of initial contact, reason for the divorce of the last/first marriage
Relations with children	Evaluation of the relations with children by those aged 60 or older, intergenerational transfers
Work	See Figure 13
Personal income	Non-operating income, operating income, financial support from relatives and friends, government subsidies, life satisfaction
Time use *(shared)*	Life, work, study and training, leisure and social activities, commuting
Leisure	Leisure activities and frequencies, way of travel, overseas experiences
Mobile phone and internet *(shared)*	Use of mobile phones, QQ, MSN, e-mails, importance of internet, frequency and places of internet use
Social relations	Help seeking, confiding troubles, self-reported social status
Subjective measurements	Values, social attitudes, achievement scale
Politics	Experiences of thefts or robbery, unjust treatments, media interests, evaluation of government work
Health	Height, weight, self-reported health, discomfort, chronic diseases, experiences of hospitalization, medical expenses, coping with diseases, satisfaction with health care, traditional Chinese medicine, physical exercises, diet, P-ADL, smoking and drinking, sleeping, memory, depression scale, main health care-giver
Basic tests	Literacy, arithmetic, physical activities
Personal information and the observations of the interviewers	Contact information, respondent, personal characteristics of respondent

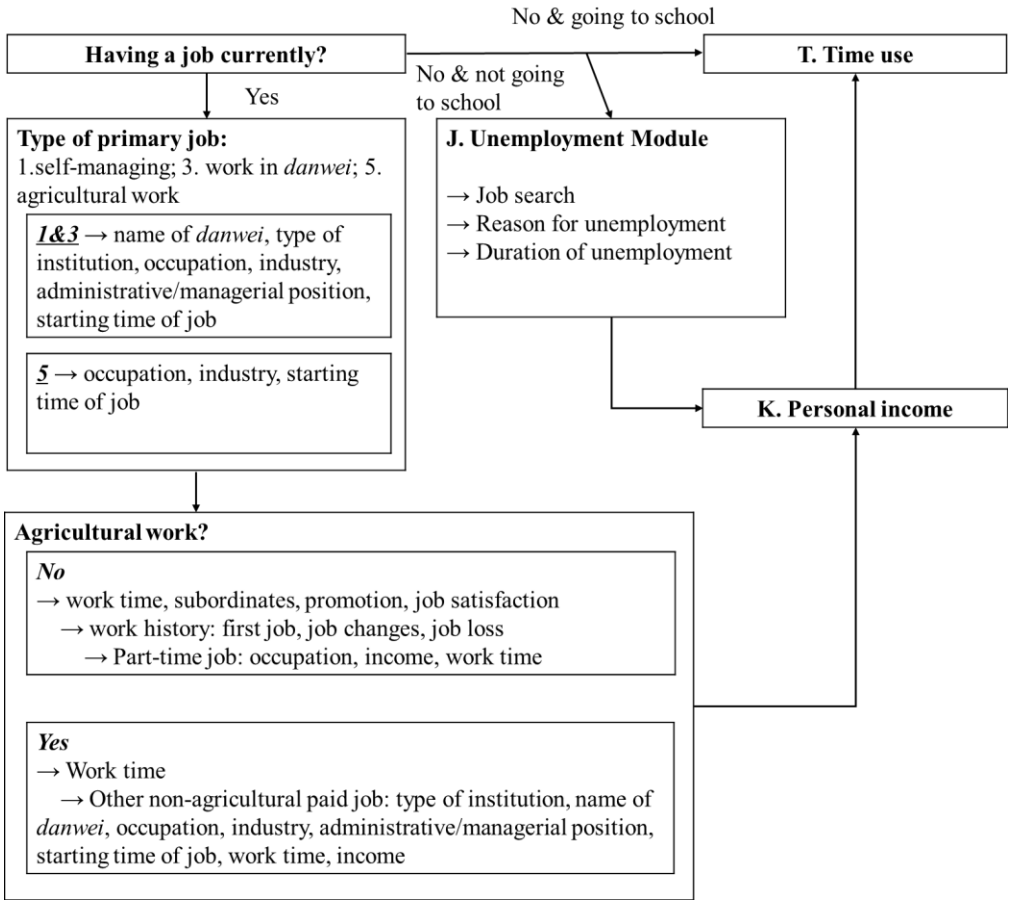


Figure 13. Flow Chart of the Module Structure of the Adult Questionnaire of CFPS 2010

4. Implementation

4.1 Pilot studies

Pilot survey is an important step prior to the baseline survey. Before the CFPS national baseline survey launched in 2010, we conducted two pilot studies in 2008 and 2009. Between May and September in 2008, a pilot survey was initiated in Beijing, Shanghai and Guangdong. This pen and pencil pilot survey focused on society, econmy, education and healthcare. The pilot survey in 2008 consists of 2,400 households, with 800 households in each province/municipality locating in 8 counties/districts. Four villages/neighborhood communities are included in each county/district, and 25 households are included in each village/ neighborhood community. CFPS adopted PPS sampling, and interviewed 2,375 households, 7,214 individuals in 24 counties/districts, 95 villages/neighborhood communities.[25]

Between May and September in 2009, the pilot follow-up survey was conducted with the sampled households in 2008. However, two types of samples were not included: households who had moved out of the village/neighborhood community, and respondents who had left their households. In the official survey, these two groups were both tracked. The sample size of the pilot study in 2009 was 1,995 households.

Unlike in 2008, the Computer Assisted Personal Interviewing (CAPI) technology was introduced in 2009. During the study, we have conducted comprehensive tests of CAPI technology, including the stability and reliability of its real-time interview management technologies, real-time technological support, and real-time data quality control.

These two pilot studies laid a solid foundation for the baseline survey in 2010. In addition, data collected in the two surveys from these three regions is available for further research.

4.2 Interviewers

For a better control of the implementation cost and time, we employed local interviewers mainly from the sample neighborhoods communitiescommunitys in the 2010 baseline survey. Each interviewer was in charge 2 neighborhoods communities, communities and in big cities, the number of interviwers should be 1.2-1.5 times more.

Interviewers were mainly recruited online. After the resume screening, phone interviews, and face-to-face interviews, a total of 453 interviewers were selected. They were divided into 14 groups to receive 6-day training at Peking University. The training courses were taught in small classes from February 22nd to August 13th in 2010, including tutorials, group practices, simulation surveys, and field interviews in real-seetings, etc. Finally 438 interviewers passed the trainings and exams and became official interviewers.

[25] The Institute of Social Science Survey (ISSS), Peking University (2009).

Table 7 below shows the basic information of interviewers in 2010. More detailed information on interviewer recruiting process and trainings for the 2010 baseline survey could be found in *The Implementation Report* (CFPS-3).

Table 7. Basic Characteristics of the Interviewers of the 2010 Baseline Survey[26] (Total Number of Interviewers: 438)

Characteristics	Category	N	%
Gender	Male	294	67.1
	Female	144	32.9
Marriage	Unmarried	265	60.5
	Married	173	39.5
Age (year)	18-19	10	2.3
	20-29	306	69.9
	30-39	101	23.1
	40+	21	4.8
Educational attainment	Graduate	11	2.5
	4-year college	198	45.2
	Junior college	153	34.9
	High school and below	76	17.4
Occupation	Corporate employee	137	31.3
	Student	109	24.9
	Public professional services	50	11.4
	Unemployed	44	10.0
	Family planning agencies	39	8.9
	Teacher	29	6.8
	Self-employed	30	6.6

4.3 Overview of 2010 Survey Implementation

The 2010 baseline survey covered a wide range of 25 provinces/municipalities/autonomous regions, including 162 counties/districts[27] and 649 communities[28]. The implementation was divided into two parts: the extensive survey during the survey season and some supplemental surveys based on the results of the former.

The survey season started in April and ended in September, 2010. During the season, we completed interviews in 600 communities[29], including 14,852 household screening questionnaires, 14,326 family member questionnaires, 14,192 family questionnaires, 32,202

[26] Technical Report: CFPS-3.
[27] The 32 streets/towns in Shanghai included 18 districts/counties.
[28] Combined villages/neighborhoods were calculated separately.
[29] Interviews are regarded as completed after the interviewers confirmed that there were 6 times of contacts and 3 refusals.

adult questionnaires, and 8,789 child questionnaires[30]. In these 600 communities, 4,224 cases failed to fill out the family members questionnaires[31]. The main reasons were as follows:[32]

(1) Inaccuracy of the sampling frame: 1,690
(2) Refuse to answer: 1,490
(3) Failure to contact after 6 attempts: 461
(4) Not eligible and screened out: 374

The supplemental interviews mainly involved the following situations. One is those communities where the interviews were not completed or did not reach the expectation of 25 households. There were 324 communities in this case, and 118 interviewers were called to complete this part of the supplemental interviews. Among these communities, those that were not interviewed included both the communities who refused to be interviewed and those that were not at a good time for interviews during the season. For example, World Expo 2010 was held during the survey season in 2010, therefore we postponed some interviews in Shanghai in order to ensure the safety in these communities. The second part involves samples with interviewers' misconduct. We needed to do the re-sampling and start the interviews all over again concerning these cases. In 2010, we found 5 serious interviewer cheating cases in total. The detailed information of these cases and their solutions can be found in *Implementation Report* (CFPS-3). In order to ensure the consistency and validity of the statistics, Computer-Assisted Personal Interviewing was applied in the supplemental interviews. The content of the questionnaires, as well as the interviewing system, was exactly the same as the first interviews in the survey season.

After the supplemental interviews, we in toal completed 635 communitycommunitys questionnaires, 15,717 household screening questionnaires, 14,960 familyroster questionnaires, 14,798 family questionnaires, 33,600 adult questionnaires and 8,990 childchild questionnaires in 2010. The sample maintenance survey in 2011 and the follow-up survey in 2012 confirmed that there are no wrong addresses and respondents, no substitute interviewers, or data fabrication in the 2010 baseline survey.

4.4 Refusal and solutions

Refusing to be interviewed causes substantial sample losses. There were more than 1,000 households who refused to be interviewed in the 2010 CFPS baseline survey.[33] The analysis found that the response rate mainly depends on the type of the community. In general, in such communities with the majority of officials or community committee staff, the upscale communities, the elderly communities, and the military communities, the respondents were more resistant to interviews and the refusal rate was relatively higher. Moreover, as the village/neighborhood committees were involved in contacting village/neighborhood residents for

[30] Technical report: CFPS-3.
[31] Technical report: CFPS-3.
[32] Technical report: CFPS-3.
[33] This is process data.

our interviews, whether the committee staffs were willing to cooperate and howand the relationship between the committees and the communities was also had a direct influence on the respondents' responsesrespondents to interviews. If the committee made less effort in the cooperation with our survey or in their management of the community, or did not have the actual administrative power over the community, the refusal rate was also relatively high.

Several solutions have been proposed to solve the refusals and other factors that lead to low rate of interview completion. First, according to the requirements in the implementation process, we would require interviewers pay more visits and ask the coordinators and other respondents who had completed the interviews to try to persuade the stubborn respondents in order to change their mind. The case could only be suspended after 3 serious refusals, and the interviewers need to inform their supervisors about the reasons of refusals and fill out a form explaining the case. Second, we would mail persuasive letters, ISSS newsletters and copies of *China Report*[34] to the respondents to gain their trust in our program. Finally, we would arrange several trials by our best interviewers, group of supervisors, and staff members from the National Family Planning Commission to help tackle the task. In this way, we have made some progress in solving the refusal problem.

4.5 Baseline survey final contact result

Tables 8-10 show the contact results at the household level in the 2010 baseline survey, and Tables 11 and 12 show the contact results at the individual level. Tables 8 and 9 show the implementation rates at household and individual levels. For detailed statistical approaches and calculation methods, please refer to Sample Contact (CFPS-5).

Table 8. Distribution of the Completion Statues of CFPS Baseline Survey Sample Units[35]

Status		Neighborhood Community		Village		Overall	
		N	%	N	%	N	%
Eligible	Interviewed (I)	5,081	63.09	9,879	82.79	14,960	74.85
	Refused (R)	348	4.32	143	1.20	491	2.46
	Incomplete due to other reasons (O)	14	0.17	25	0.21	39	0.20
	Not contacted (NC)	17	0.21	36	0.30	53	0.27
Not eligible (NE)		652	8.10	812	6.80	1,464	7.33
Unsure of eligibility (UE)		1,941	24.10	1,038	8.70	2,979	14.91
Total		8,053	100	11,933	100	19,986	100

[34] The ISSS will prepare a descriptive report, known as the *China Report*, when each wave of CFPS is completed. The *China Report* has been published by Peking University Press twice prior to the 2010 baseline survey, based on the 2008 and 2009 pilot surveys, respectively.
[35] Technical Report: CFPS-5.

Table 9. Distribution of Non-eligible Sample Units of the CFPS Baseline Survey[36]

Type	Neighborhood Community		Village		Overall	
	N	%	N	%	N	%
Wrong address	62	9.51	23	2.83	85	5.81
Non-residence	97	14.88	37	4.56	134	9.15
Vacant house	412	63.19	628	77.34	1,040	71.04
Failure in household screening	81	12.42	124	15.27	205	14.00
Total	652	100	812	100	1,464	100

Table 10. Distribution of the Sample Units with Unsure Eligibility in CFPS Baseline Survey[37]

Type	Neighborhood Community		Village		Overall	
	N	%	N	%	N	%
Address not contacted	7	0.36	5	0.48	12	0.40
Address correct, unable to contact the residents	662	34.11	484	46.63	1,146	38.47
Refusal to be interviewed	1,194	61.51	430	41.43	1,624	54.51
Cannot conduct household screening due to other reasons	78	4.02	119	11.46	197	6.61
Total	1,941	100	1,038	100	2,979	100

Table 11. Distribution of the Completion Statues of CFPS Baseline Survey Individual Sample Units, by Urban and Rural)[38]

Status	Neighborhood Community		Village		Overall	
	N	%	N	%	N	%
Not contacted	508	3.06	1,163	2.87	1,671	2.92
Interviewed	12,793	77.03	29,797	73.49	42,590	74.52
Refusal	1,752	10.55	2,533	6.25	4,285	7.50
Incomplete due to other reasons	539	3.25	1,535	3.79	2,074	3.63
Not eligible	1,015	6.11	5,520	13.61	6,535	11.43
Total	16,607	100	40,548	100	57,155	100

[36] Technical Report: CFPS-5.
[37] Technical Report: CFPS-5.
[38] Technical Report: CFPS-5.

Table 12. Distribution of the Sample individual with Ineligibility in CFPS Baseline Survey [39]

Reasons of absence	Current Residence Within CFPS Region		Current Residence Outside CFPS Region		Overall	
	N	%	N	%	N	%
Out for school	126	1.80	972	13.89	1,098	15.69
Out for work	294	4.20	5,135	73.36	5,429	77.56
Monk	17	0.24	59	0.84	76	1.09
Visiting families and friends	28	0.40	246	3.51	274	3.91
Serving in a sentence					19	0.27
Serving in the army					75	1.07
Abroad					29	0.41
Total	465	6.64	5,931	84.73	7,000	100

Table 13. Implementation rate of the CFPS baseline survey sample unit (%)[40]

Type	Formula[41]	Resident Committee	Village Committee	Overall
Response rate	RR3=I/(I+R+NC+O+UE)	69.35	89.16	81.25
Cumulative response rate	$RRcum=RR_{\text{resident screening}*}RR_{\text{family member}}$	69.35	89.16	81.25
Cooperation rate	COOP1=I/(I+R+O)	93.35	98.33	96.58
Contact rate	CON2=(I+R+O)/(I+R+NC+O+eUE)	74.29	90.68	84.13
Refusal rate	REF2=R/(I+R+NC+O+eUE)	4.75	1.29	2.67

Table 14. Implementation rate of the CFPS baseline survey individual unit (%)[42]

Implementation Type	Community Type		Age range		Overall
	Resident Committee	Village Committee	Adults	Children	
RR5=I/(I+R+NC+O)	82.05	85.07	82.52	90.67	84.14
COOP1=I/(I+R+O)	84.81	87.99	85.69	92.34	87.01
CON3=(I+R+O)/(I+R+NC+O)	96.74	96.68	96.31	98.29	96.70
REF3=R/(I+R+NC+O)	11.24	7.23	9.05	6.08	8.47

[39] Technical Report: CFPS-5.
[40] Technical report: CFPS-5.
[41] "e" represents the percentage of total eligible samples in the eligibility screening
[42] Technical report: CFPS-5.

4.6 Sample maintenance

CFPS made detailed sample maintenance plans to prevent sample losses and ensure effective long-term follow-up surveys. The maintenance work was done in 2010. The first round of maintenance started in September 2010, right after the survey season, targeting at all the sampled households and neighborhood communitycommunity, including both the interviewed and non-interviewed. We sent out mails, including thank-you letters for the interviewed households, for those who refused to be interviewed and for the village/neighborhood community. The second round of maintenance was implemented during the Spring Festival, targeting at the 14,767 households interviewed in the 2010 baseline survey. During this period, the Institute conducted the supplemental interviews of CFPS and surveys on the satisfaction of health care reform with some of these households. For those households, we did the "free-ride" on-site maintenance.[43] For the rest 97 communities, we made phone calls to those who had telephones and paid visits to those who did not have telephones or if the phone maintenance did not work. On such basis, we sent Spring Festival greeting cards and annual newsletters to the households with valid addresses by post. For more information on the maintenance plan and the process of the sample maintenance, or the results of statistical analysis of the maintenance outcome, please refer to *Sample Maintenance* (CFPS-18).

[43] As it turned out later, a total of 1,226 households from 59 villages/neighborhoods were not successfully maintained in the "free-ride" way. They were combined with the 97 independent villages/neighborhoods to receive the telephone/on-site sample maintenance.

5. Survey Quality Control

5.1 Quality Control Measures and Technologies

Strict measures of the quality control were applied in the 2010 CFPS baseline survey to ensure the quality of the data. For the factors that may affect the data quality, such as the improper designs of the questionnaires, inaccurate terminal-stage sampling, irregular behaviors of the interviewers, mistakes in data collection, and compilation processes, we applied different methods of monitoring and intervention, including telephone check, field check, audio record check, interview reviews, statistical analyses, and so on.

For example, if interviewers used substitutes for the sampled households/individuals or interviewed the wrong households/individuals, the representativeness of the sample would be affected and thus cause trouble for data users. Therefore, the quality control team applied different measures before and after the survey was done. To prevent arbitrary substitutions and interviewing the wrong households, the sampling staff or the cadres of the neighborhoodscommunity would deliver the letters for the households to the correct addresses before the interviews. Then the residents in these households would call the Institute to report the required information. The quality control team compared this information with information collected by the interviewers to ensure consistency. After laugching the interviews, they would also confirm the accuracy of household sampling by field check: the ing. Tfield checking team would document the number of the eligible households on the address and provide feedbacks. The filed checking team is also responsible for documenting interview refusals and non-contacts. In the case of arbitrary substitutions and interviewing the wrong individuals, the team mainly used field check, telephone check, and audio record check to perform post-hoc control the quality. In addition, if interviews were done with wrong households or individuals, re-interviews would be conducted with the correct ones. weW had put much emphasis on the behavioral code of the interviewers during their training sessions.

Moreover, regarding systematic biases caused by improper designs of questionnaires, we did statistical analysis of the questionnaire data and paradata weekly to identify the sources of errors. We would revise the questionnaires accordingly and update the interviewing system to control the quality. For inaccuracies in the terminal-stage sampling, field checks were used. For irregular behaviors of the interviewers, such leading questions and avoiding the complicated question modules by abusing logical skips, telephone check, audio record ,checkand paradata analysis of interview time and of non-responses of the questions could help control the data quality.

With the CAPI system, the likelihood of systematic biases in data collection and compilation was much lower than in the traditional surveys with paper and pencil. However, biases could not be completely avoided during the coding process of open-ended questions. Three coders were asked to code the same questions using systematic coding tables. There were also errors in the data compilation process, which could be controlled via cyclic inspections by multiple people.

5.2 Quality Control Strategies

Our quality control is comprehensive. First, all the variables in each questionnaire and paradata should be checked using statistical methods regularly (every 7 days). Second, all the interviewers should go through all the checking methods. Third, for each interviewer, data from each questionnaire need to be checked. Fourth, each type of contact results by each interviewer also needs to be checked.

Our quality control has a substantial coverage. First, 60% of the addresses of incomplete interviews received on-site filed checking. Second, the households whose questionnaires for families or individuals were completed should be checked via audio record, phone calls, on-site, or interview reviews by the percentage of 15%, 25%, 15% and 5%, respectively.

Regarding the order of quality control, statistical analysis of data was always used for overall data varification. When the sample size of each type of questionnaires reached 30, the data should be merged every 7 days to check for systematic errors. Audio record check was the preferred measure to be used in the overall check, which started on the second day after receiving the collected data from interviewers and continued to the third day after the end of the survey season. The phone check was preferred for the cases without audio records or cases that failed the audio record checks, which started on the second day after receiving collected data. The field check was mainly used to examine the accuracy of the terminal-stage sampling, which started within 20 days after receiving the collected data. The interview review was mainly used for the problems found in previous forms of checks, such as too short or too long interview time. The whole process of quality control has emphasized on timely responses.

5.3 Proportions and Results of Quality Check[44]

A total of 28% of all the households interviewed received audio record checks successfully, covering 16% of all the completed questionnaires.

A total of 19% of all the households interviewed received telephone checks successfully.

During the survey season, a total of 25% of all the households received field checks. In addition, during the periods between December 2010 and February 2011, and July to November 2011, the Institute conducted another two rounds of field checks on a part of the sample and the entire sample, respectively.

For interview reviews, 3% of the valid cases were involved.[45]

The statistical analyses cover the interview time, non-response rate, outliers, internal consistency reliabilities of the attitude scales, and so on. Reports were submitted to the quality supervision department and the department of survey implementation on a weekly basis.

The results of quality control show that there were no interviews of wrong households among all the households checked via telephone, in field and record. A total of 81 adult

[44] Results come from the Technical Report: CFPS-4.
[45] In the mid-late period of the survey, the method of interview reviews was replaced by the statistical analysis of the time used in the interviews due to its high cost and low efficiency.

questionnaires were not answered by themselves, with 21 interviewers involved. Among them, 59 questionnaires were due to the misconduct of one single interviewer; 7 questionnaires were answered by other family members when the respondents were not at home; for the rest 15 questionnaires, either the respondents were unable to answer questions or others cut in to answer the questions. For the child questionnaires, 22 questionnaires that should be answered independently by children themselves were completed by others, with 20 interviewers involved. Among those cases, 7 questionnaires were answered by parents. In addition to these cases, we also discovered several serious cases of misconduct in field checks, which were documented in Chapter 4 of this manual.

Most interviewers strictly complied with the behavioral code of interviews.The quality control suggested that the short interview durations in some interviews were due to interviewers' misbehaviors such as speeding up and skipping the questions. We constructed a "sample of short interviews" that included the interviews with very short interview time and the interviews where the questions answered in a very short time account for more than 50% of all the questions monitored. This sample contains 1,051 cases in total (including questionnaires for families and individuals) and involves the work of 50 interviewers. Among all questionnaires that were randomly checked, 7,914 have at least 1 question omitted.[46] We took specific actions on this problem during the survey and made significant improvement.

The quality control of the 2010 CFPS not only was effective in watching for the data quality but also gained much experience in the measures and strategies of quality control. The detailed design and the quality control process can be found in *Quality Supervision Report* (CFPS-4).

[46] The quality of the interview could not be simply judged by the problem of leaving out questions. Among the questions omitted, despite some that were left out on purpose, the vast majority of them were obvious according to the observations or the information that the interviewers already acquired, e.g. when the interviewer saw the respondent calling on the mobile phone, he would directly answer yes in question "Do you have a mobile phone" without asking.

6. Data Sets and Variables

6.1 General Introduction to Data Sets

The CFPS 2010 data consists of data sets based on the community questionnaires, the family roster questionnaires, the family questionnaires, the adult questionnaires, and the child questionnaires.

As introduced in the chapter on sampling, there are 6 sampling frames representing 6 subpopulations. In the data set released, we marked different subpopulations with the indicator variable "subpopulation". The values from 1 to 6 of the "subpopulation" represent Shanghai, Liaoning, Henan, Gansu, Guangdong, other provinces/municipalitiesrespectively. Also, we added a dichotomous indicator "subsample". The value 1 of the "subsample" refers to the resampled national sample.

The complete national sample included all the CFPS data, composed of the 6 subsamples representing the 6 subpopulations. After weighting, the complete national sample represents the national population. The resampled national sample was constructed by resampling the five "big provinces" proportionally to the "small provinces." The resampled national sample is directly representative on the national level (i.e., without weighting). For details on weights, please refer to Chapter 8.

Table 15. Number of Variables and Sample Size of CFPS 2010 Data Sets

	N of Vars	Sample Size						
		Full sample	Resampled sample	Shanghai	Henan	Gansu	Liaoning	Guangdong
Community	222	635	417	58	64	65	63	64
Family roster	340	57,155	36,964	4,329	6,491	6,874	4,652	6,423
Family	624	14,798	9,661	1,405	1,506	1,537	1,478	1,394
Adult	1,490	33,600	21,812	3,162	3,732	3,704	3,129	3,070
Child	961	8,990	5,944	360	1,273	1,213	529	1,115

Note: See Codebook for detailed information of the variables.

In the following, we focus on introducing our data processing work and providing explanations of some special or difficult variables as well as notable issues in data usage.

6.2 Data Cleaning

6.2.1 Cleaning of the Family Relations Data set

The main content of the family relations data set come from the three tables T1, T2, and T3. As mentioned above, these three tables were used in the questionnaires to collect social-demographic information of family members and their relations, which would provide important information and help for the data cleaning in the later period.

The data cleaning of the entire family relations data set was divided into several stages. In the first stage, the basic cleaning focused on correcting errors in the implementation process, such as dealing with wrong information input by interviewers and invalid questionnaires (e.g., cheating questionnaires, and repeated questionnaires), correcting the wrong sample codes of the households, etc. This part of data cleaning was mainly based on the information feedback from the implementation team in the field.

In the second stage, the basic match cleaning focused on the correspondence between the individuals in the family relations data set and the individuals in the adult and child data sets. According to the designs of CFPS, the family members with their individual codes starting with the number "1" in the family relations data set need to answer the individual questionnaires. Therefore, if the contact result showed that they had completed their questionnaires, we need to find the corresponding individual questionnaires in the individual data sets in this stage. At the same time, we should confirm that the questionnaire found in the individual data sets matches the family member himself/herself—there must be strict correspondences. We checked the correspondence through some key variables such as names and the dates of birth and corrected mistakes in the confirmation process.

The third stage is the deep cleaning of the family relations data set. After the first two stages, we had already done the matching work in the following two aspects:[47] First, we matched all the family members and the immediate relatives in T1 and T3 tables according to the relation index in T2 Table. We found that there were some logic errors or questionable facts in some samples, such as female fathers, male mothers, or the situations where the husband and wife did not know each other, or the age gaps between parents and children were too small or too large, etc. Second, we matched the family relations data set and the individual data sets and also found some logic errors or questionable facts, such as situations where the total number of the children reported by the respondent's spouse was different from the number by the respondent, or the spouse was no longer alive in the family based on family relations data but the marital status in the individual data sets was "married" or "cohabitation", etc.

For the data cleaning in the third stage, we first migrated to a new system version and solved some errors in the codes of family members caused by the computer program. All other errors were corrected manually based on reliable information sources such as the 2010 family relations data set and the 2010 individual data sets, the 2011 family relations data set and the 2011 individual data sets. If there were no reliable references, we normally defined the absolute logic errors as "missing", e.g., wrong gender for the father, same gender for spouses. For other questionable details that were not absolutely wrong, e.g., the small age gap between parents and children, we left them unchanged. For the detailed process and methods of manual data cleaning, please refer to *Data Cleaning of the Family Relations Data Set* (CFPS-7).

6.2.2 Cleaning of Other Data Sets

The cleaning of the other data sets was also divided into several stages.

[47] See detailed matching method in the technical report: CFPS-6.

Data cleaning in the first stage mainly focused on data sets of individual questionnaires, which was done at the same time with the deep cleaning of the family relations data set for cross-validations. On the one hand, while cleaning the family relations data set and checking the correspondence between the information in these two data sets, we found some mistakes in the individual data sets, e.g., the big gap between the age of marriage claimed by the spouses, the spouse was no longer alive in the family relations data set whereas the marital status in the individual data sets was "married" or "cohabitation", etc. If the problem was with the individual data sets after careful inspections, we would correct the errors in this data set. On the other hand, there were some logic errors and questionable details in the individual data sets itself, e.g., the date of divorce was prior to the date of marriage, the date of marriage was prior to the date of birth, the age of marriage was less than 16, etc. In these cases, we relied on the information from the family relations data set to make corrections. The data cleaning principle of the individual data sets is consistant with the data cleaning principle for family data set.We would deal with the mistakes if there were reliable information sources. If not, we defined those absolute errors as "missing" and left those uncertain situations unchanged. Moreover, for important variables, instead of directly correcting the original values, we constructed a group of new "best variables" that summarize the most reasonable information after data cleaning. See detailed introduction to the "best variables" in the section of "Composite variables" below.

In the second stage, we did the cleaning in all other data sets besides the family relations data set, lasting from the end of the first stage to the final release of the data. With all the careful checks over all the variables, the data team deleted redundant variables, supplemented some of the missing variables, and corrected wrong variable names, labels and values, as well as some obviously wrong values.[48] However, although we paid great attention to some extreme values that were likely to be unreasonable, in most cases we accepted the original values unless there was particularly reliable evidence for correction. Data users are allowed to make their own judgments.

6.2.3 Field Work Collection and Data Editing

With the efforts of the data team, most of the problems in the data sets had already been solved. However, there were still some data from a small proportion of households/individuals, which could not be corrected due to the lack of reliable information or evidence. Of those data, we tried to supplement or confirm some key information, for which we made the following efforts.

(1) Return Visits

For households with logic errors in the family relation data set, which we were unbable to clean due to the lack of information, we conducted return visits in 2012. These visits mainly aimed for fixing the following three types of errors. The first type was that we could make certain corrections due to the lack of information. For example, as for the unreasonable age gap

[48] For example, one household chose the value of "-8" in the variable fe3 "Do you participate in or manage any non-agricultural industries", but there was valid data in the non-agricultural module indicating the answer to this question. Therefore we changed the value of the variable fe3 to "1" (yes).

between parents and children (less than 15 or more than 50), the reasons of this error could be: ① it was indeed the fact; ② the ages of the parents were wrong; ③ the ages of the children were wrong; ④ they were not their biological children; ⑤ the other family members' information was mistakenly entried instead of the children's. If we could not acquire accurate information in the data cleaning process, we put these families into our list of return visits. The second type was the blood relationship between parents and children. According to the original design of CFPS, the parental relationship should refer to the blood relationship (including adoption) and exclude fosterage. But in field operations, this rule wasn't strictly obeyed, which caused the different standards of defining parental relationship in the survey. Therefore, we listed all the reconstructed families in our return visits and screened the type of their children so that the blood relatives and fostered children could be accurately separated. The third type was that there were individuals who had no family relations. They either made up a household with themselves alone, which was rather rare in reality if they did not have any immediate relatives; or had no relations with other family members although they belonged to the households in permanent residence. These individuals' information was marked as questionable, for they might not be eligible as family members or some of their family relations were omitted in the survey. To regain their family relations and clarify the definition of family members in the baseline survey, we decided to visit households with this type of individuals again in the year 2012.

Data problems vary from family to family, which require interviewers active probing. Therefore, a standard questionnaire and interviewing procedures cannot be used in the return visits. Instead, we conducted paper and pencil interviews with case-by-case questionnaire designs, each of which included specific questions and the specific way of asking and answering questions for specific families in order to confirm or retrieve information. The return visits went simultaneously with our 2012 survey. For the households on the list of return visits, the interviewers needed to complete the case-by-case questionnaires specially designed for the specific families after finishing all the interviews in the survey of 2012. At the end of the survey in 2012, the information we collected in the return visits was used to correct and update the 2010 data.For the first and third types of errors, we supplement and correct the information in the data sets. For the second type of errors, we created a new variable (bio_cN). This variable represents whethere the person, who takes turn to answer the questionnaire and the corresponding child are blood related or not. Among which, value "1" represents there is a bollod relationship between the household head and the child, "0" represents not blood related, '-8" represents not applicable, and "-9" represents data missing.[49] The bio_cN variable can help researchers understand the blood and adoption relationships in reconstructed families, which also explain why there might be the reasons why parents and children do not recognize each other and why the age gaps between the parents and the children are small.

For more information on the return visits, please refer to *Data Cleaning of the Family Relations Data Set* (CFPS-7)

(2) Repeated Questions, Confirmations, and Supplemental Questions

[49] For the person who takes turn and does not have remarriages, no values are given to the corresponding child in terms of the child's bio variables

In addition to revisits, we implemented questions in the survey of year 2012 to reconfirm information gained in survey of year 2010. The reconfirmation mainly focuses on important variables, which were hard to correct due to lack of information, and key variables, which were influenced by improper quesionniare design or problematic implementations in 2010. Detailed information is listed below:

① Gender and the date of birth of the respondent.

② Marital status in 2010 and important time information in his/her marital history, e.g., date of the marriage, spouse's date of birth, and time of divorce, etc.

③ For the respondents who were divorced or widowed in the interviews in 2010, we asked about the education level of their last spouse since the T table design in 2010 did not collect this information. We recollected it in the questionnaires in 2012 considering that it would be useful for researchers who are interested in family and marriage studies.

④ Supplementary information on education history. The designs of CFPS in 2010 and 2011 could only collect the education experiences of the adults over 16. As for children under 16, they were only asked about their highest educational level (for those who were not at school) or the current educational level (for those who were at school). Their detailed education information was not acquired. Therefore, we supplemented this part of information on education by interviewing the respondents who did not answer questions on their education experiences in our surveys in 2010 and 2011. Moreover, we reconfirmed the current state and level of education in the questionnaires of CFPS 2012.

⑤ Questions on the education level when respondents left school. The question on educational level in the questionnaires of CFPS 2010 was "What is the highest level education you have completed?" This question actually did not consider the last educational level of those who had quitted school without graduation, which led to the underestimation of their educational level. Thus, the question was modified as "What educational level were you in when you left school?" in our questionnaires for all the respondents who had left school without graduation, in order to make an accurate estimation of their educational level.

⑥ The information of the respondents' parents, including their date of birth, their occupations, education levels and political statuses. In the field survey of CFPS 2010, we did not collect such information of the parents who had passed away. Their information was collected in 2012 to make up for the data in Table T3.

After finishing the 2012 survey, we updated the data of CFPS 2010 based on information collected above.[50] We hereon remind the data users to refer to the updated data of CFPS 2010 and the data of the 2012 survey to adjust and supplement data set if needed.

[50] We will release multiple relevant technical reports (e.g. renew method, data evaluation, suggestions of variable usages) to interpretate data updates. See current relevant technical reports: CFPS-21 and CFPS-22.

6.3 Composite Variables

For the convenience of data users, we created some composite variables that are commonly used in research. Below is a brief introduction to their meanings and constructions. Detailed rules of constructions can be found in the corresponding technical report.

6.3.1 Educational Level

The educational level and years of education of individuals are two of the common variables used in social research. The latter is usually imputed on the basis of the former. In CFPS, the variable "education level" had multiple information sources (see Table 16). Due to differences in family membership, age interval, and status of schooling, multiple respondents have provided multiple answers about educational levels in the family relations data set and individual data sets, which might be confusing and inconvenient for data users.

Table 16. Information Sources of the Variable "Educational level"

		Items
Family relations	Table T1	B4 (all the family members living together) the highest education degree acquired
	Table T3	D5 (immediate relativeimmediate relatives living together) the highest education degree acquired
Children: self-answered (ages 10-15)	Going to school	R1 What educational level are you in R401 What grade are you in? .....
	Not going to school	H4 In what educational level did you master's degree? .....
Children: answered by others (ages 10-15)	Ages 3-5	F3 Is the child currently in primary school? F303 In which year did the child start primary school?
	Ages 6-15	F3 Is the child at school currently? F301 What educational level is the child currently in? F302 What grade is the child in currently? F303 In which year did the child start to go to school? .....
Adults	Going to school	R1 What educational level are you in currently? R401 What grade are you in currently? .....
	All adults	C1 What is the highest education degree you have acquired (with graduation) up till now? C105/C205/C305/C405/C503/C603/C703 How many yeas have you studied? C107/C207/C307/C407/C505/C605/C705 Have you got the education degree or the gradation diploma? .....

For data users' convenience, we created the following composite variables based on these different sources of information:

edu: educational level (crude)

eduL: years of education (crude)

educ: educational level (detailed)

eduy: years of education (detailed)

The variable "edu" is the education level with 8 crude categories: illiterate/semi-illiterate, primary school, junior middle school, senior middle school, junior college, college, master's degree, and doctoral degree. This variable is mainly based on the information answered by the respondents themselves and adjusted by the answers by others. The variable "eduL" is the years of education they received imputed based on the variable "edu".

All types of schools were asked in CFPS, e.g., junior middle schools include ordinary junior middle schools, junior middle schools for adults, vocational junior middle schools; senior middle schools consist of ordinary senior middle schools, senior middle schools for adults, ordinary polytechnic schools, polytechnic schools for adults, vocational senior middle schools, and technical schools. The variable "edu2" is a more detailed measure for educational level with 16 categories. And the variable "eduy" is the years of education more accurately calculated based on "edu2" and other factors such as the type of the schools, the grade, and graduation statuses, etc.

The calculation methods of the educational level and the years of education are shown in Tables 17 and 18 below.[51]

Table 17. Educational Level and Years of Education, Crude Versions

Code	Educational level (edu)	Years of education (eduL)
1	Illiterate/Semi-illiterate	0
2	Primary school	6
3	Junior middle school	9
4	Senior middle school	12
5	Junior college	15
6	College	16
7	Master's degree	19
8	Doctoral degree	22

[51] Technical report: CFPS-12.

Table 18. Educational Level and the Years of Education, Detailed Versions

Code	Educational level (educ)	Years of education (eduy)	
		Graduated	Not graduated (including still attanding)
1	Illiterate/Semi-illiterate	0	0
2	Primary schools for adults/Verbal class	3	1.5
3	Ordinary primary school	6	3
4	Junior middle school for adults	9	7.5
5	Vocational junior middle school	9	7.5
6	Ordinary junior middle school	9	7.5
7	Polytechnic school for adults	12	10.5
8	Senior middle school for adults	12	10.5
9	Ordinary polytechnic school/Vocational senior middle school/Technical school	12	10.5
10	Ordinary senior middle school	12	10.5
11	Other junior college/for adults/nline	15	13.5
12	ordinary junior college	15	13.5
13	Other college/for adults/online	15	13.5
14	ordinary college	16	14
15	Master's degree (including working professionals)	19	17.5
16	Doctoral degree (including working professionals)	22	20.5

Similarly, we created the variables on the education levels and the years of education for fathers, mothers, and spouses. See Table 19 for the variable names and labels. For the detailed methods of creating the composite variables on educational level and preliminary results of statistical analysis, please refer to *Composite Variables (II): Educational Level and Depression Scale* (CFPS-12), *Evaluation and Analysis of Differences in Self-reporting and Proxy on Education Level* (CFPS-21), and *Composite Variables (IV): Parents' Social Status* (CFPS-22)

Table 19. Educational Level and Years of Education for Father, Mother, and Spouse

Name	Label	Name	Label	Name	Label
fedu	Father's educational level (crude)	medu	Mother's educational level (crude)	sedu	Spouse's educational level (crude)
feduL	Father's years of education (crude)	meduL	Mother's years of education (crude)	seduL	Spouse's years of education (crude)
feduc	Father's educational level (detailed)	meduc	Mother's educational level (detailed)	seduc	Spouse's educational level (detailed)
feduy	Father's years of education (detailed)	meduy	Mother's years of education (detailed)	seduy	Spouse's years of education (detailed)

6.3.2 Depression

CFPS applied the same scale to measure the mental state of individuals in the adult questionnaires and children (Table 20). We performed factor analysis and used the factor score to indicate the level of depression, which is named "fdepression". The variable "depression" is the additive index score based on the 6 questions. For detailed information of these two composite variables, please refer to *Composite Variables (II): Educational Level and Depression Scale* (CFPS-12).

Table 20. The Depression Scale of CFPS 2010

Adults Question No.	Children Question No.	Items
Q6	N4	Please select the following statements about the mental state according to your situation in the recent month: 1. almost every day 2. often 3. half the time 4. sometimes 5. never
Q601	N401	The frequency of feeling upset, depressed, unable to be excited whatever you do in the recent month?
Q602	N402	The frequency of feeling stressed out in the recent month?
Q603	N403	The frequency of feeling anxious, unable to keep calm in the recent month?
Q604	N404	The frequency of feeling hopeless in the recent month?
Q605	N405	The frequency of feeling difficult doing everything in the recent month?
Q606	N406	The frequency of feeling meaningless in life in the recent month?

6.3.3 Cognitive Ability

The 2010 CFPS baseline survey applied a group of verbal and mathematical problems to test and evaluate the cognitive ability of all respondents who needed to complete the questionnaires by themselves (i.e., children aged 10-15 and all adults).

The verbal problems (X2 in the self-answered child questionnaires, X1 in the adult questionnaires) had 8 groups of problems of approximately the same difficulty level. In each group, there were 34 Chinese characters, listed from the easiest to the hardest. In the interviews, the computer would randomly select one group for each respondent to answer. According to different education levels in T1 table, the respondents started to answer the verbal questions at different starting points. Those with "1-2" (primary school and below) started from the first character; those with "3" (junior middle school) started from the ninth; and those with "4-8" (senior middle school and above) started from the 21st. The characters were presented in cards and the respondents were asked to read them aloud. If they could not read 3 characters in a row or the questions went to the 34th character, the test ends.

The mathematical problems include addition, subtraction, multiplication, division, exponents, logarithms, trigonometric functions, sequence, permutation and combination, etc. All the problems were divided into four groups of similar difficulty level. The computer would randomly select one group for the respondents to answer. Each group had 24 problems, also listing from the easiest to the hardest. Similar to the verbal problems, the starting points were based on the educational level in the T1 table: those with "1-2" (primary school and below) started from the first; those with "3" (junior middle school) started from the 5th; andthose with "4-8" (senior middle school and above) started from the 13th.

Considering that these two groups of problems might be re-used in follow-up surveys, we did not release the problems to the public for the consideration of the validity for future tests. We calculated the scores for these two groups of problems, named "wordtest" and "mathtest", respectively. We assigned the scores according to the question number of the most difficult problem that the respondent had answered correctly. If the respondent did not give any correct answer, the question number of the problem prior to where he/she started would be the final score.

While using these two variables, please note that they are not standardized and not adjusted to age or educational level. Users may need to further process these variables according to specific need in research.

For the detailed methods of creating these variables and the results of preliminary statistical analysis of these variables, please refer to *Composite Variables (I): Verbal and Mathematical Tests* (CFPS-11).

6.3.4 Best Variables

In the process of data cleaning, for some variables that needed to be corrected, we created new variables based on the original ones instead of correcting them directly. Both kinds of variables were made available to data users. As the newly created variables have the most reasonable values which take into consideration various information sources, we call them "best variables". The rule of naming these variables was to put "_best" at the end of the names of the original variables. For data users, please note that "best variables", which are the most adequate values based on answers provided, multiple information sources, and logic relations, are not necessarily the correct variables. Researchers can use either the original answers or the "best variables."

The "best variables" of the survey in 2010 were: qa1y_best, qe605y_best, qe606y_best, and qe1_best. These four "best variables" are all in the adult data set.

The variable "qa1y" describes the year of birth of the adults. In the 2010 CFPS baseline survey, we could acquire this information for members in Table T1 from three sources (see Table 21): 1) the information in the questionnaires on family members answered by others, 2) the information provided by the respondents themselves, and 3) spouses' answers after spousal matching, which has the assumption that the current spouse (or first spouse) answered the individual questionnaire as well.

Table 21. Information Sources of the Year of Birth

	Questionnaire	Question No.	Question
Source 1	Questionnaires for family members	B1	What is the date of birth of "family members?"
Source 2	Adult questionnaires	A1	What is your date of birth?
Source 3	Adult questionnaires	E606/E302/ E211	What is the year of birth of your first spouse/current partner/current spouse?

Through statistical analysis, we found that the information from these three sources were not always consistent, which might lead to inconvenience for the data users. Therefore, we modified this variable both manually and with the help of computer program. After checking the logic relations between the year of birth and several other life events, we eliminated the unreasonable values and gave the best values of the year of birth, named "qa1y_best". We did not delete the original values with the "best" ones for two reasons. First, it was because they were only the most reasonable values after our careful considerations but we were not sure if they were 100% correct. Second, the age of the respondent was directly connected with the type of the questionnaires he/she had answered. Either their answers were correct or not, the questionnaires of individuals were automatically created according to the year of birth they claimed. The original values were the evidence for the questionnaires, which is the reason why we must keep them in the data set. Otherwise, it might confuse the data users in terms of the rules of creating the questionnaires.

The variable "qe1" is about the current marital status in the questionnaire for adults. We also acquired this information in the questionnaires for family members and there were some inconsistencies between these two information sources. Moreover, we found some logic errors during the data cleaning of the family relations data set mentioned above, e.g., the marital status was "unmarried" while the respondent had a spouse in the T2 table; or the state was "divorced" or "widowed" while there was a living spouse in the T2 table, etc. Some of these errors resulted from wrong answers of the marital status. In this case, we created the variable "qe1_best" as the most reasonable marital status for data users. Again, we did not directly change the original values, for the marital status would affect the question orders in the marriage module. If the values of the variable "qe1" were changed, it would lead to unnecessary confusions for data users.

In the module of marriage history in CFPS questionnaires, we used a retrospective way to collect respondents' information on the important events in their marriages, enabling researchers to study marriage and its relationship with other events. However, the retrospective way as well as some sensitive questions might pose a challenge to the accuracy of the data collected. As the respondents may not remember things precisely the way they were, or they would hide away some facts on purpose, in the data cleaning process, we found many logic errors (e.g., the time of the second marriage was prior to the time of the divorce of the first marriage), inconsistencies of different information sources (e.g., the couple gave different answers on the time of marriage or on the spouse's year of birth), answers that violate the common sense (e.g., the age of the first marriage was below 16), etc. We chose the most important two variables: "qu605y", the year of the first marriage and "qe606y", the year of birth of the first spouse to create corresponding best

variable. After eliminating the errors in the family relations data set in data cleaning, we referred to different information sources, e.g., the age interval of proper marriage, the year of birth of the first child, etc., and kept the most reasonable values. Thus we created the "best variables" "qe605y_best" and "qe606y_best". We preserved the original values with the "best" ones. First, we could not guarantee that the "best" values were 100% correct—they were just the most reasonable values in our view. Data users can select variables based on their own judgments. Second, we believe that some inconsistencies are worth studying, e.g., the different answers on the time of marriage given by the couple, the different memories of the spouse's year of birth, etc.

For the detailed information on the method of creating these four "best variables", please refer to *Composite Variables (III): "Best Variables"for Age and Marriage* (CFPS-13).

6.3.5 Income

In the 2010 CFPS baseline survey, for rural families, we only asked about the income from selling agricultural products in farming, forestation, herding, side-line production, and fishing; we did not ask about the values of the produce that were consumed by the rural families themselves. As the rural families in China would consume a considerable amount of agricultural products by themselves, the data collected in CFPS cannot give an accurate estimation of the total income of agricultural production of these rural families. The omission of the portion for self-consumption causes underestimation of the actual income from agricultural production, especially in less marketized regions and rural families in poverty.

To correct for this, we designed an adjustment scheme for agricultural income of rural families based on information obtained in the quesionniare, and generated two variables "inc_agri" and "net_agri", which represent the adjusted total income and the adjusted net income from the families' agricultural production, respectively. The adjustments mainly based on information on the families' production output, sales volume, sales income, and net income (K6 and K7 in the family questionnaires). We calculated the amount of products for family consumption by comparing the total production output and the total sales volume. Then we converted this portion to income according to the market prices. The total and net incomes of the family's agricultural production were imputed by adding this converted amount of income to the total and net sales incomes. For detailed information on the adjustments of agricultural production incomes, see *Adjustments of Rural Family Income* (CFPS-14).

Table 22 compares the means, standard deviations, and the medians of the net income of agricultural production before and after the adjustment.[52] Adjustments were done for all families working on agricultural production in 2009. It is notable that the mean of the net income per household increased by 2,469.8 RMB and the median increased by 2,275.5 RMB after modification.

[52] Technical report: CFPS-14.

Table 22. Net Income of Family Agricultural Production, before and after Adjustment

	Mean (RMB)	Standard Deviation	Median	N of Household
Original	4,889.0	12,796.6	2,105.0	7,586[53]
Modified	7,358.7	14,044.3	4,380.0	7,586

Due to the complexity of economic conditions, there were many questions on personal and family income in the 2010 CFPS baseline survey. For data users' convenience, we created a group of composite variables base on personal and family incomes. The relevant variables and labels are listed below in Table 23.

Table 23. Composite Variables on Personal and Family Incomes

Name	Label	Name	Label
Income	Personal income	faminc_net_old	Unadjusted net family income
firm	Non-agricultural income	faminc_old	Unadjusted total family income
finc	Wage income	faminc	Adjusted total family income
fpropert y	Property income	faminc_net	Adjusted net family income
welfare	Transfer income	indinc	Adjusted total family income per capita
felse	Other income	indinc_net	Adjusted net family income per capita
foperate	Adjusted total business income (including agricultural production and non-agricultural businesses)	foperate_net	Adjusted net business income (including agricultural production and non-agricultural businesses)

The calculations of the variables listed in Table 23 are as follows.

- Personal income: first we used the personal income from the questionnaires as the values; if the information was missing, we used the mean of the income zone instead; if this mean was still missing or less than 100, we replaced it with the total income calculated by all the items.[54]
- Total income: the income with the cost of agricultural production included.
- Net income: the income without the cost of agricultural production.
- Total income or net income per capita: the average income with the total income or net

[53] There were 7,798 households that reported to have worked on agricultural production. Here the 7,586 households were the cases that had no missing values both before and after adjustments. This sample includes urban families working on agricultural production.

[54] The items calculated in the total income included wages, awards, annual bonus, welfare items, the income of the second job or part-time jobs, the income of other work, pensions, self-employed income, economic aids from relatives and friends, economic aids from the villagers'/neighborhood committees, the allowances from the government or work places.

income divided by the family size. Here the family size is the number of family members living together shown in the T1 table.

- Total/net family income: the total calculation of the five parts of income, i.e., wage income, total/net business income, property income, transfer income, and other income.
- Wage income: wages, awards, allowances, income of working out of town, and bonuses for the individual.
- Business income: agricultural income (income from farming, forestation, herding, side-line production and fishing) and non-agricultural income (for other non-agricultural production).
- Property income: rent of land or other means of production, house rent, other rent income, income from selling properties. As for financial property income such as interests on savings, stocks, funds, and bonds, since CFPS only asked respondents the savings the families held and the corresponding market values at the end of 2009 but the income might be consumed long before the surveys, we excluded this part of income from property income for the current year.
- Transfer income: government allowances, pensions, and other economic aids from the government such as basic living allowances.
- Other income: gifts from relatives and friends, andother income claimed by the respondent families.

Apart from these composite variables, we also cleaned the missing values and extreme values in the income and expenditure part of the survey. For the calculation standards of each type of the income and detailed analysis, please refer to *Income and Expenditure Data* (CFPS-15).

6.3.6 Conversion of Occupational Codes

The coding system (GB/T 6565-2009) of Chinese Standard Classification of Occupations (COSO) was applied in the 2010 CFPS baseline survey. For data users' convenience, we converted the occupational codes to several other schemes and created some composite variables related to occupations.

(1) The corresponding coding system of occupations in CPFS is International Standard Classification of Occupation (ISCO-88). These variables have "_isco" after the original variable names.
(2) There were two sets of socioeconomic indices based on ISCO-88, namely the International Socio-Economic Index of Occupational Status (ISEI) and Treiman's Standard International Occupational Prestige Scale (Treiman's SIOPS). The naming rule of these two sets of variables is to add "_isei" or "_siops" after the original names.
(3) The Erikson and Goldthorpe's Class Categories were applied to the adult data on occupations. The naming rule is to add "egp" after the original names.

Table 24. Examples of the Names of the Occupation Variables

	Father's primary occupation (B5 in questionnaires for family members)	Sibling 1' primary occupation (B309 in adult questionnaire)	Post-hoc codes for respondent's occupation (G307 in adult questionnaire)
CFPS occupation variable	tb5_code_a_f	qb309_occu_1	qg307code
ISCO-88 code	tb5_isco_a_f	qb309_isco_1	qg307isco
ISEI score	tb5_isei_a_f	qb309_isei_1	qg307isei
SIOPS score	tb5_siops_a_f	qb309_siops_1	qg307siops
EGP category	——	——	qg307egp

In the meantime, we provided the Stata commands of the conversions from CSCO codes to ISCO88, ISEI, Treiman's SIOPS, and EGP.

For detailed construction methods and content of the composite variables on occupation, please refer to *Conversion of Occupational Codes and Construction of Socioeconomic Indices* (CFPS-10).

6.4 Occupation Codes

Two schemes were applied in the 2010 CFPS baseline survey to code the respondents' occupations: field coding by interviewers and post-hoc coding by the data team.

The field coding was used in questions G307, G308, H405, and H406. With the help of the CAPI system, the interviewers directly coded the occupations and industries of the respondents according to the coding dictionary of CFPS occupations and industries.[55] The interviewers went from the broad categories to detailed categories (4-digit codes) of occupations and industries. The interface of the coding system is shown in Figure 14.

Apart from above four questions, other questions on occupations and industries are all open-ended, which were coded manually by trained data workers after the survey. In the coding process, we controlled the data quality with Two-way Independent Verification with Adjudication. In the first round, two data workers coded the occupational information of each respondent separately. If their results matched, it was then confirmed as the final code; if inconsistent, a third experienced data worker would code these items in the second round. The second code would be accepted if it was consistent with either of the two in the first round. If not, we would ask expert researchers to make decisions based on the codes by data workers, the field codes, and other auxiliary information.

[55] The CAPI method is still in the trial stage. Given the complexity of occupation classifications in China, it is difficult to confirm the category of occupations. Therefore, interviewers also recorded detailed occupational information (G306).

Table 25. Questions and Auxiliary Items on Occupations and Industries of CFPS 2010

Questionnaire	Question No. (variable name)	Item	Type of Question
Family roster	B5 (tb5)	What is the main occupation of "family member"?	Open-ended
Family roster	D6 (td6)	What is the main occupation of "immediate relatives living separately"?	Open-ended
Adult	B309 (qb309）	What is the main occupation of "sibling"?	Open-ended
Adult	G303 (qg303)	What institution are you currently working at?	Multiple choice
Adult	G304 (qg304)	What is the name of your working place?	Open-ended
Adult	G305 (qg305)	The institution you are working at belongs to?	Multiple choice
Adult	G306 (qg306)	Your occupation?	Open-ended
Adult	G307 (qg307)	Which category does your occupation belong to?	CAPI-assisted field coding
Adult	G308 (qg308)	Which industry does your occupation belong to?	CAPI-assisted field coding
Adult	G601 (qg601)	What is your first occupation?	Open-ended
Adult	G701（qg701）	What is your second occupation?	Open-ended
Adult	H404（qh404）	What is your non-agricultural occupation?	Open-ended
Adult	H405（qh405）	What is the category of your non-agricultural occupation?	CAPI-assisted field coding
Adult	H406（qh406）	What industry does your non-agricultural belong to?	CAPI-assisted field coding
Children	J2（wj2）	What is the primiary duty (the one that takes most time) in your formal work?	Open-ended

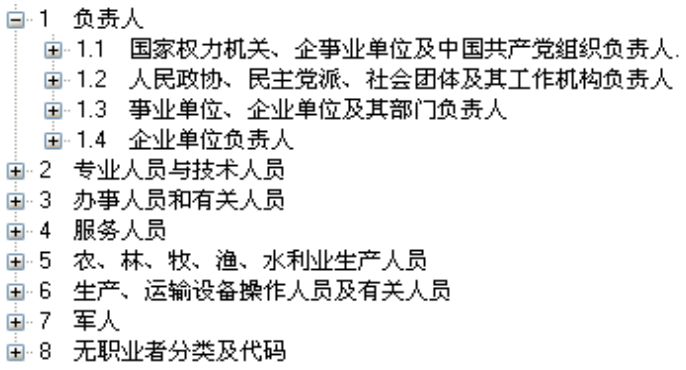


Figure 14. Computer Interface of the Field Coding System of CFPS 2010[56]

[56] Technical report: CFPS-18.

Due to substantial differences between the field codes and the post-hoc codes, we adopted the latter for questions G307, G308, H405, and H406.

The latest version of the national standard classification of occupations GB/T 6565-2009 was not released in the design stage of the baseline survey questionnaires. Therefore, the 2010 survey used a modified version of GB/T 6565-1999, drawing on the classification standards of occupations and industries in the Chinese General Social Survey (CGSS). This scheme classifies occupations into 8 categories, including 595 occupational codes. The classification of industries was adopted from the standards of the National Bureau of Statistics, which contains 20 categories.

In the post-hoc coding, we applied the latest version of national standard classification of occupations, i.e., GB/T 6565-2009. The categorization and ordering were entirely adopted, yet the codes were re-labeled.

For detailed technical methods and quality evaluation of occupation codes, see *Occupation and Industry Codes of CFPS 2010* (CFPS-8).

6.5 Confidentiality Issues

Personal information is strictly protected in CFPS. We made the following decisions for confidentiality.

(1) The respondent's name is not released. Instead, we provide personal ID for data users.
(2) The exact date of birth is not released. Year and month of birth are available.
(3) We only release the address information at the provincial level. For the address information at district/county and village/neighborhood levels, we only offer numerical codes for nominal differentiation.
(4) In question G1 in the adult questionnaire, options 5 and 6 were merged as "others". The question G103 is not released.
(5) The question M706 in the adult questionnaire is not released.

6.6 Miscellaneous

Table 26 provides a brief summary of some variables that require special attentions in the use of data.

Table 26. Special Instructions on the Use of Selected Variables of CFPS 2010

Content	Questionnaire	Question No. (Var Name)	Brief Instructions
Interactions and relationships between respondents over 60 and their children	Adult	F1-F3(qf1-qf3)	Children's information based on T tables. May contain errors due to matching of family relations. Corrections made in the 2012 survey. Users please refer to the 2012 data.[57]
Addresses information, e.g., places of birth and *hukou* registration	Adult	A102、A201 (qa102acode, qa201acode)	Address information at the provincial level available. For the address information at district/county and village/neighborhood levels, numerical codes provided for nominal differentiation.
Career expectation	Adult	S8, S801	No information collected due to an error in the CAPI.
Career expectation	Child	M601(wm601), D101(wd101)	Coded. See CFPS-9: *Career Expectation Codes*
Launage in daily communication	Adult & Child	Adult D2 (qd2), Child K2 (wk2), Share S3 (ks3)	Coded. See CFPS-20: *Dialect Codes*.
Metric, e.g., How far is it from the nearest senior middle school to your home? ____m/*li*/km.	Family	A6(fa6)	Standardized units available in variable labels.
Coding of multiple choices, e.g., What food have you eaten in the recent month?	Child	L5(wl5)	Instead of 0/1 coding for each option, a set of variables are created (i.e., w15_s_1 to w15_s_9). The value of w15_s_1 records the first option chosen by the respondent, and the value of w15_2_2 records the second option chosen, and so on. If the respondent chose 3 options, the values from ws15_s_4 on would all be -8; same as other multiple-choice questions
"Others" options, e.g., What is the reason	Adult	J101(qj101)	In semi- open-ended questions, verbal information specified in

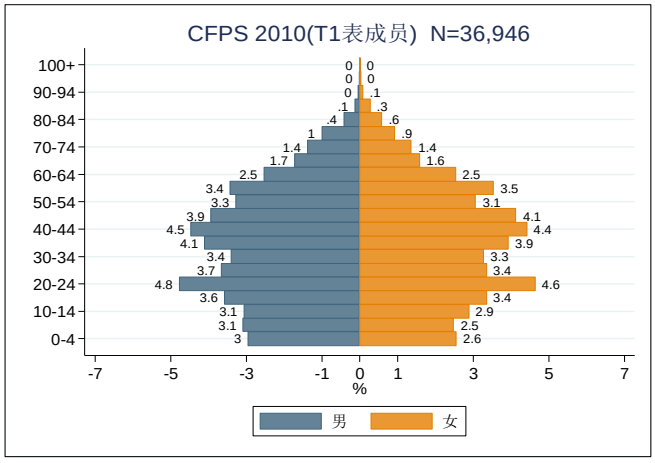
[57] Please notice that although we asked the questions based on the information of the children in the survey 2010, the questions we asked again were about the specific situations in 2012 rather than the recall on the information of 2010.

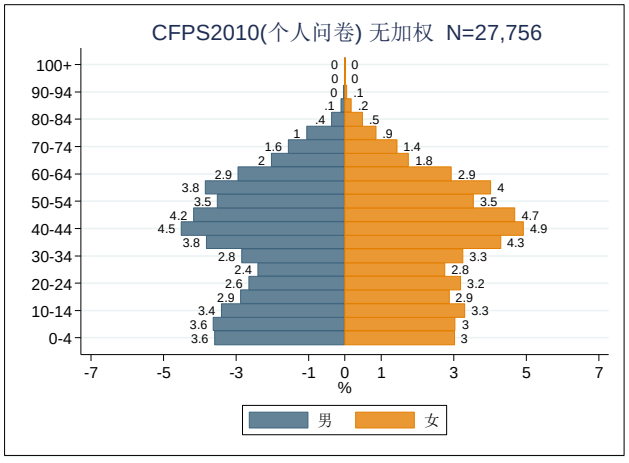
that you are currently unemployed? 77. Others [Please specify]____			option “77” is not released; same as other semi-open-ended questions.

7. CFPS 2010 Baseline Survey Preliminary Findings and Evaluations[58]

7.1 Age-Sex Distributions

Figure 15 presents the age-sex population pyramids based on the resampled national sample of the CFPS 2010 baseline survey, tabular data of the 2010 national census,[59] and the 2010 Chinese General Social Survey. In the population pyramids, we break ages down to 5-year groups across the range from age 0 to age 100 and above, and present sex-specific relative sizes of each age group. For CFPS, we present the age-sex structures of both the T1 table members (all co-residing household members) and the respondents who have completed the individual questionnaires (adults and children). According to the design, all co-residing family members were target subjects of the survey. However, in the actual survey, only some of the family members answered the questionnaires due to physical absences or refusals. Therefore, the age-sex pyramid based on all T1 table members reflects the characteristics of the sampled national population, and the one based on all completed individual questionnaires reflects the characteristics of the respondents who completed those questionnaires. Similarly, for CGSS 2010 we also give the age-sex structures of all the family members in the interviewed households and of the actual respondents—the former represents the population while the latter represents the age-sex characteristics of all survey respondents.





[58] All results in this chapter come from the resampled national sample.
[59] The data comes from Tables of the 2010 Sixth National Population Census "T3-01 National Population by Age and Sex".

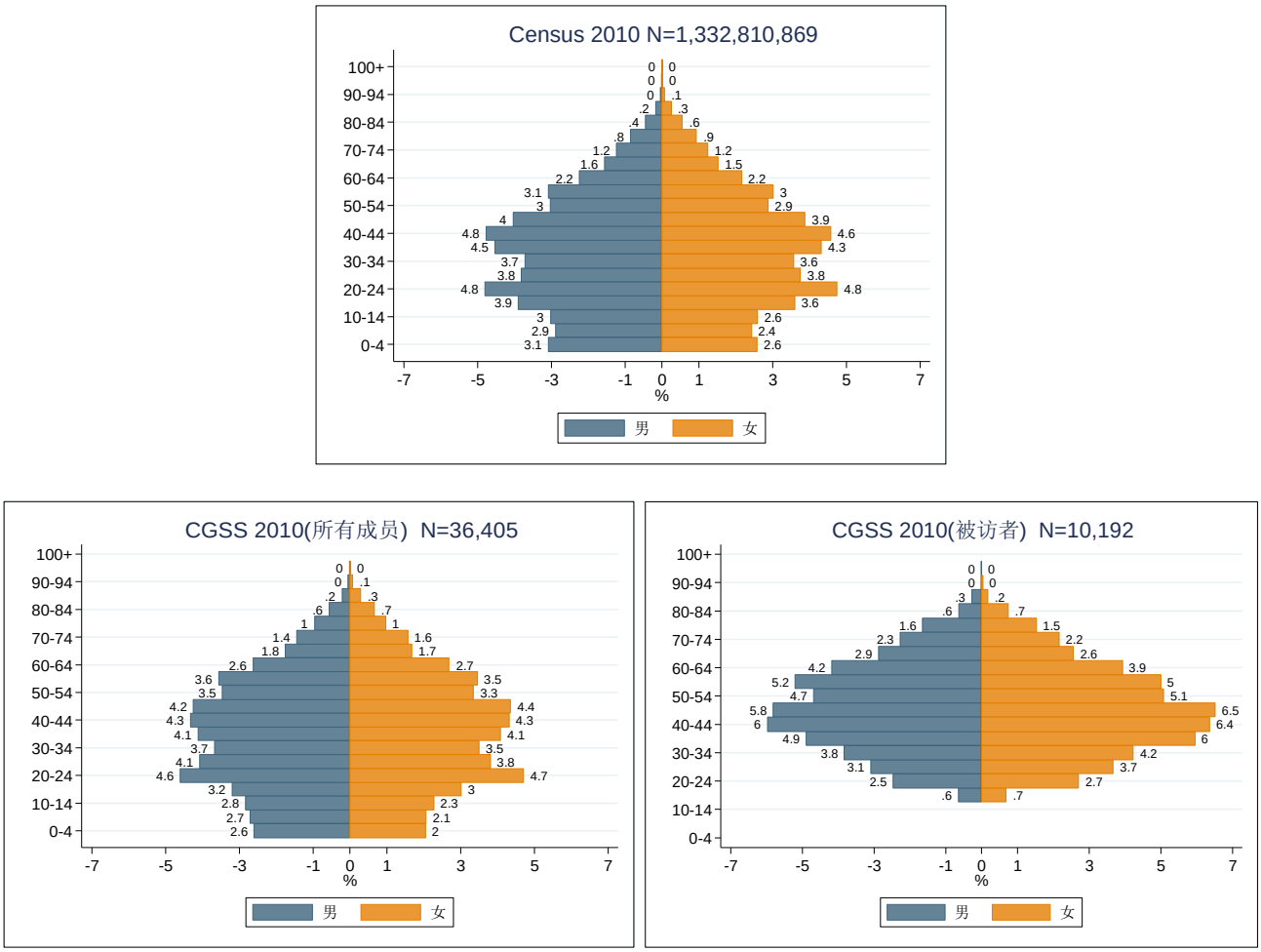
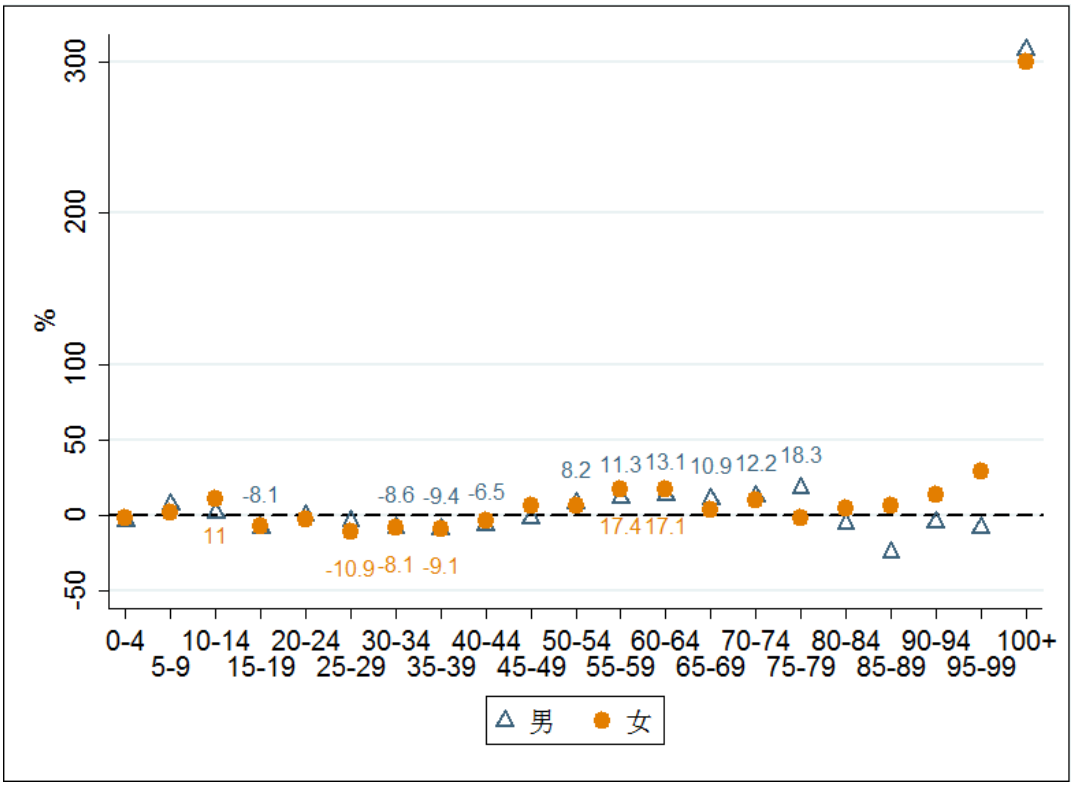


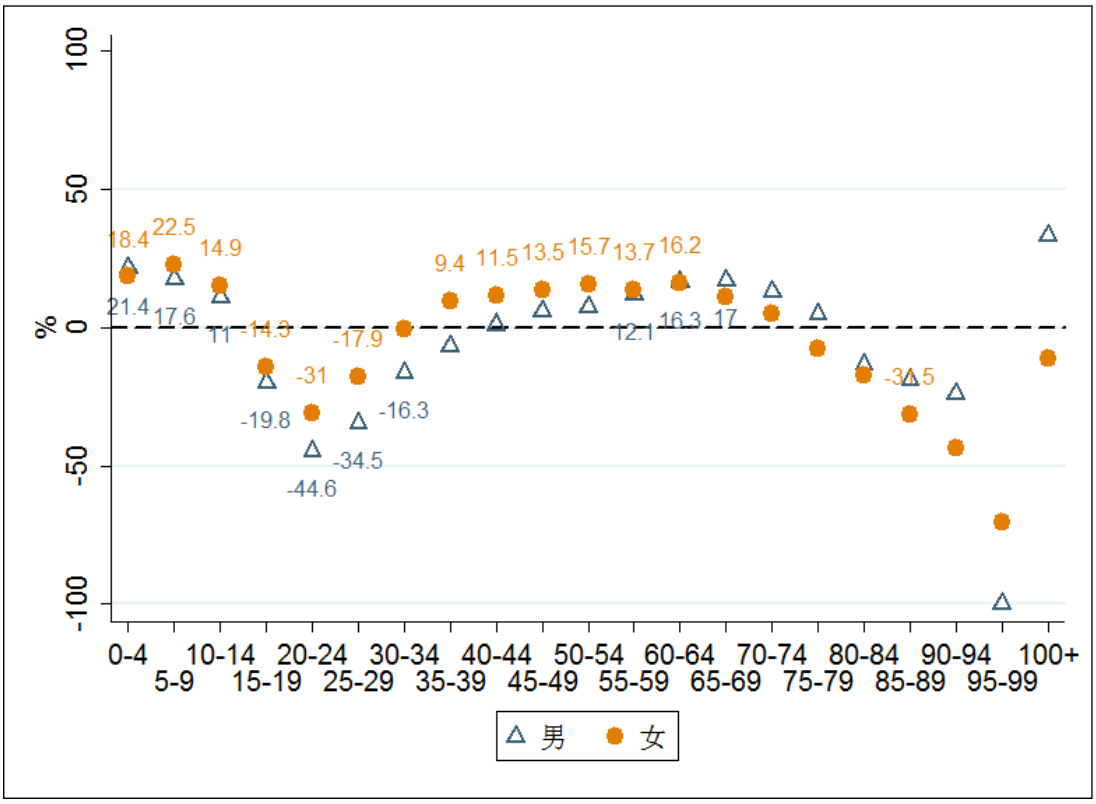
Figure 15. Population Pyramids of CFPS 2010, Census 2010, and CGSS 2010

We compared the age-sex structures of both CFPS T1 members and all members of interviewed households in CGSS with the population structure as in the 2010 national census. We found that the shape of the pyramids of these two surveys is rather similar to census data: the 20-24 and 40-44 age groups are the largest, and the young and old groups are the smallest in size. Specific to each age group, the CFPS data is closer to the census data. We applied log-rate models to test the differences between the two survey data structures and the census population structure, with the population age-sex frequency distribution as the exposure. This method assumed that the census data was the population and the data from CFPS and CGSS were probability samples drawn from the population. If the sample age-sex structures were consistent with the population structure, the sampling probability of each individual should be the same. We included gender, age and their interactions into the full model and compared it to the null model, which produced a Chi-square value for model comparison. The Chi-square value of comparison between the T1 table members of CFPS and the census is 253.70 (*df*=41, p=0.0000), and the Chi-square value for CGSS is 529.85 (*df*=41, p=0.0000). The data compositions of both surveys are significantly different from the population age-sex structure, but CFPS is closer to the census, which has a smaller Chi-square value.



Note: Points with value labels are significantly different from the census data at 0.01 level. Those without value labels are not significantly different.

Figure 16. Differences between CFPS 2010 T1 Members and the 2010 National Census



Note: Points with value labels are significantly different from the census data at 0.01 level. Those without value labels are not significantly different.

Figure 17. Differences between CFPS 2010 Individual Respondents and T1 Members

Figure 16 shows the differences between the T1 members of CFPS and the census population in the frequency distribution of age-sex groups. The triangles and the dots in the figure correspond to male and female groups, respectively. If there are no differences, the value should be 0. Values greater than 0 mean that the relative sizes of the age-sex groups in CFPS are larger than the corresponding groups in the census, while values smaller than 0 mean that the proportions are smaller. We marked out the significant differences at the 0.01 level in the figure. We can see that the sampling probability of females aged 10-14 is relatively higher, that of males aged 15-19 is lower, those of males and females aged 25-44 are lower, and those of males aged 50-79 and females aged 55-64 are higher. In general, the age-sex structure of the T1 members of CFPS is consistent with the 2010 national census.

In terms of the age-sex structure of the CGSS respondents (Figure 15), there is no representation of the age group 0-14 and the proportion of the age group 40-60 was relatively higher, as only one family member aged over 18 in each selected household was interviewed in CGSS. Compared with CGSS, the respondents of CFPS individual questionnaires had a wider age-sex distribution, with respondents from all age groups. The age groups over age 30 in CFPS data are closer to the census, but the proportions of younger groups (aged 15-19, 20-24, and 24-29) are much lower, since these individual are more likely to go out of town for work or study. We examined the difference in the age-sex structure between all T1 members and all respondents of individual questionnaires. We took the frequency of each age-sex group of individual

respondents as the dependent variable and the frequency of each age-sex groups of the T1 members as the exposure. Comparing the full log-rate and the null model, we reached a Chi-square value of 1,059.8 (df = 41, p = 0.0000), which meant that the age-sex structure of the actual respondents is significantly different from the structure of the ideal sample according to the survey design. Figure 17 shows the differences between all individual respondents and all T1 members of CFPS: the actually interviewed proportions of both male and female aged 0-14 are relatively higher, those of the females aged 15-19 and of males aged 15-34 are lower, those of females aged 35-64 and of males aged 55-69 are higher, and those of females aged 85-89 are lower. Such differences in the age-sex structure of the actual respondents would affect the accuracy of statistical inference, which can be solved by weighting on unit nonresponse. For the calculation method and instructions of the weight, see *Weight Calculation* (CFPS-17).

7.2 Family Size and Household Types

In the CFPS resampled national sample, the average family size calculated from T1 table is 3.8. Conditional on co-residence, the number is 3.3. The CFPS family size is larger than that based on the 2010 national census. T test shows that the difference is significant at 0.001 level. Broken down to rural and urban areas, the CFPS family sizes are larger than those calculated from the census, which is true with the case of all T1 members as well as the case with the presence T1 members. Figure 18 also lists the average family sizes in CGSS, which are very close to CFPS both in national averages and by rural and urban areas.

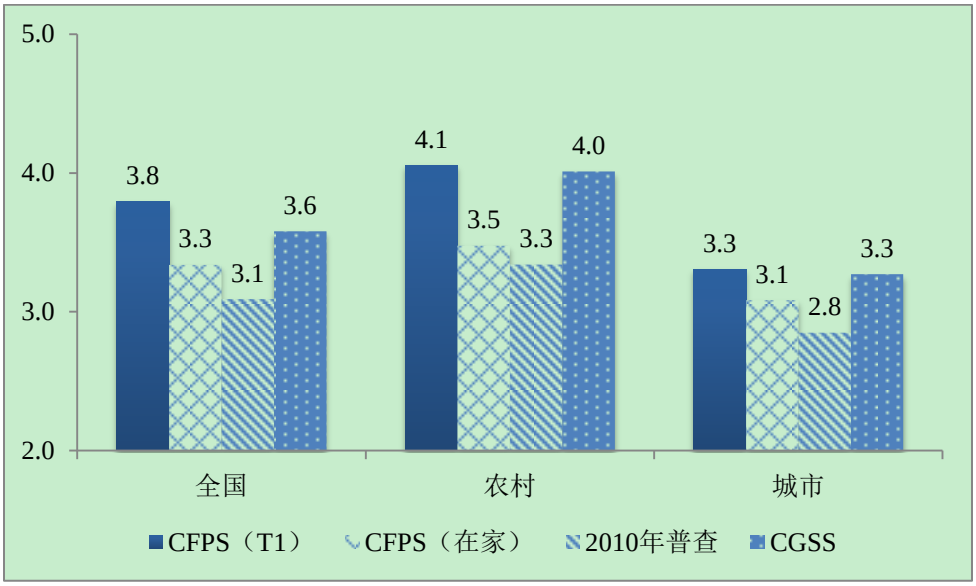


Figure 18. Average Household Size, by Rural and Urban

In the CFPS resampled national sample, based on the family members of T1 table, one-generation households account for 20.2% all households, two-generations households account for 48.7%, and three-or-more-generation households account for 31.2%. If the calculation is restricted to the members present at home, the proportions are 29.3%, 42.8%, and 28.0%,

respectively. The proportion of the three-or-more-generation households is larger than that in the census while the proportion of one-generation households is lower. The Chi-square test is significant at 0.001 level. In comparison, the family types in CGSS are closer to the CFPS results.

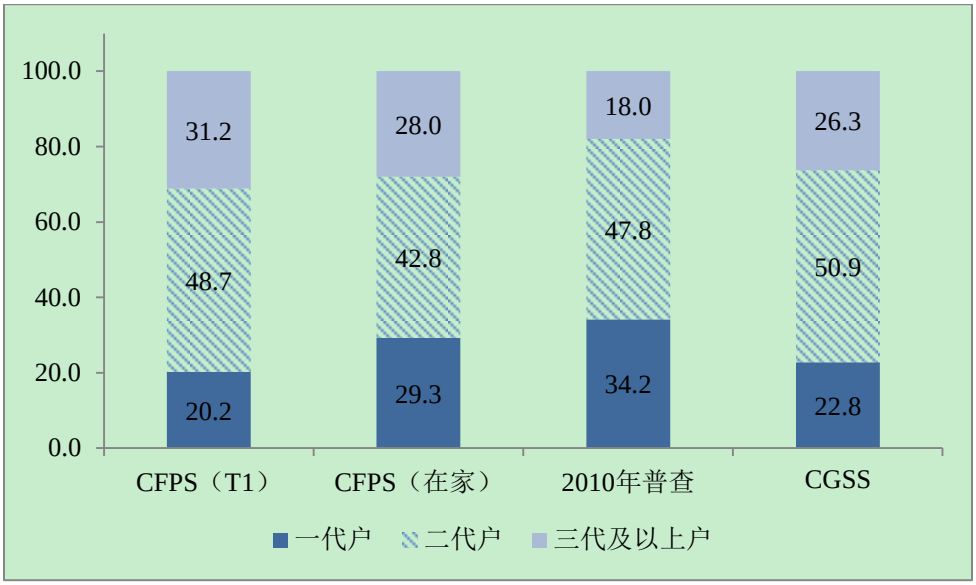


Figure 19. Household Types at the National Level

7.3 Family Income

Table 27 shows mean family incomes and Gini coefficients based on three different national samples: CGSS, CFPS, and CHFS (China Family Financial Survey). The results from CGSS and CFPS are closer. The Gini coefficient for urban families is higher in CGSS. However, in terms of mean family income and Gini coefficient, CHFS gives substantially higher estimates than CGSS and CFPS. The mean urban family income of CHFS is 87,071 *yuan*, which is nearly twice the amount of the CFPS estimate. The Gini coefficients for both rural families and urban families are greater than 0.65 in CHFS.

Table 27. Sample Size, Mean Family Income & Gini Coefficient: CGSS, CFPS, and CHFS

		CGSS2010	CFPS2010 (adjusted)	CHFS2011 (24 provinces)
Rural	Sample size	5,313	5,883	4,162
	Mean income	22,125.8	28,826.4	32,285.8
	Gini	0.495	0.498	0.675
Urban	Sample size	3,849	3,248	3,619
	Mean income	53,494.0	44,917.6	87,071.4
	Gini	0.535	0.470	0.655

Note: The CHFS 2011 data includes 25 provinces, which include Qinghai but not Fujian compared to CFPS 2010. The Qinghai data was excluded to improve comparability. In addition, zero or negative family incomes were excluded.

Tables 28 and 29 show the distributions of rural and urban family incomes based on these three samples. There are two versions of the CFPS income data: the unadjusted and the adjusted. The unadjusted distribution is calculated on the basis of the original data from the survey and the adjusted uses the total rural family income that includes the value of self-consumed products.[60] As shown in these two tables, the income distributions of CFPS and CGSS are very close for both rural and urban families. However, the CHFS data presents a rather polarized distribution. For the lower income (below 25th percentile), the proportion given by CHFS is much lower than that in the other two samples, especially for rural families. On the other hand, the proportion of higher income families (above 75th percentile) is much higher, especially for urban families.

Table 28. Distribution of Total Rural Family Income: CFPS, CGSS, and CHFS (*yuan*)

Percentile	CGSS2010	CFPS2010 (unadjusted)	CFPS2010 (adjusted)	CHFS2011 (24 provinces)
5th	2000	1940	2300	720
10th	3240	3300	4210	1750
25th	8000	9000	10000	5220
50th	15000	18000	20000	13250
75th	28880	32000	34205	32249
90th	45000	53600	56121	61025
95th	60000	74000	77628	93000

Note: The CHFS 2011 data include 25 provinces nationwide, with Qinghai added to replace Fujian. Therefore the data of Qinghai Province was excluded from the samples in order to make these data more comparable. In addition, family samples with income lower than 0 are excluded here.

[60] See detailed adjustment methods in *Adjustments of Rural Family Income* (CFPS-14).

Table 29. Distribution of Total Urban Family Income: CFPS, CGSS, and CHFS (*yuan*)

Percentile	CGSS2010	CFPS2010 (unadjusted)	CFPS2010 (adjusted)	CHFS2011 (24 provinces)
5th	6000	5000	5000	1700
10th	10000	9640	10000	5170
25th	20000	17800	18000	19001
50th	30000	30200	30300	38450
75th	55000	55000	55000	78000
90th	100000	86000	86100	169000
95th	150000	120000	120000	262500

Note: The CHFS 2011 data include 25 provinces, which include Qinghai but not Fujian compared to CFPS 2010. The Qinghai data was excluded to improve comparability. In addition, zero or negative family incomes were excluded.

Tables 30 and 31 describe the percentile distributions of rural/urban family income based on these three different survey data. The results of CFPS and CGSS are very close below the 50th percentile while CHFS proportions are relatively lower. Particularly, in the CHFS rural sample, the total income of the families at the lower half accounts for only 10% of the total income of all the families. For families at 5th and 10th percentiles, the results of CHFS are also significantly lower than the other two surveys—as shown in both Tables 28 and 29. For families at higher percentiles (75th, 90th, 95th), the rural sample results of CGSS and CFPS are very close; and CGSS urban sample shows a higher level of clustering along the distribution. However, CHFS has the highest level of clustering along the distribution in both rural and urban areas. Particularly, for the rural sample, CHFS shows that the top 5% income group has 43.1% of the total income.

Table 30. Percentile Distributions of Rural Family Income: CGSS, CFPS, and CHFS (%)

Percentile	CGSS 2010	CFPS 2010 (unadjusted)	CFPS 2010 (adjusted)	CHFS 2011 (24 provinces)
Below 5th	0.3	0.2	0.2	0.0
Below 10th	0.9	0.7	0.8	0.2
Below 25th	5.5	4.1	4.6	2.0
Below 50th	18.0	16.3	17.5	10.0
Above 75th	59.6	61.2	59.7	72.5
Above 90th	36.0	38.5	37.1	53.4
Above 95th	24.1	27.1	25.9	43.1

Table 31. Percentile Distributions of Urban Family Income: CGSS, CFPS, and CHFS (%)

Percentile	CGSS 2010	CFPS 2010 (unadjusted)	CFPS 2010 (adjusted)	CHFS 2011 (24 provinces)
Below 5th	0.3	0.3	0.3	0.1
Below 10th	1.4	1.1	1.4	0.4
Below 25th	7.7	5.7	6.2	3.6
Below 50th	16.9	19.0	19.1	14.0

Above 75th	63.0	57.6	57.6	65.6
Above 90th	39.2	35.0	35.0	43.9
Above 95th	30.1	22.8	22.9	33.2

7.4 Urban-Rural Distribution

We made a comparison of the rural-urban distribution between CFPS, 2010 census, and CHFS. As CGSS used stratified sampling for rural and urban areas and oversampled urban areas, the data from CGSS is not suitable for comparison with CFPS in this case.

Table 32 shows that the distributions of the T1 members and individual respondents of CFPS are very close based on the rural-urban division standard defined by the National Bureau of Statistics, though the Chi-square test of the two-way contingency table shows statistically significant differences ($\chi^2(1) = 5.0157$, $p = 0.025$). In contrast to the nearly even rural-urban distribution (the proportion of respondents in urban areas is higher than those in rural areas by 0.6%) of the 2010 national census, the CFPS data show that the proportion of rural residents is nearly 10% higher than the urban residents. Both the log-rate analysis ("rural" as the reference group and census as the exposure) ($\chi^2(1) = 462.83$, $p = 0.000$) and the Chi-square test of the two-way contingency table ($\chi^2(1) = 462.03$, $p = 0.000$) show that there is a significant difference in the rural-urban distribution between CFPS and the census. This may be due to sampling differences or the response rates. Also, we classified "towns" into the urban areas without considering that they may also have rural areas within its administrative or geographical boundaries. Due to lack of lower levels of census data, we are currently unable to find out the true reasons. Compared with CFPS, the rural-urban distribution of CHFS is closer to the census data, although significant difference still exists ($\chi^2(1) \approx 150000$，$p = 0.000$).

In terms of the division by the type of village/neighborhood community, the frequency distributions of T1 members and individual respondents of CFPS are very close, although the Chi-square test of the two-way contingency table is significant ($\chi^2(1) = 10.8773$，$p = 0.001$). Approximately 70% of the respondents in CFPS were living in communities under the administration of committeesvillage.

The rural/urban division by *hukou* in the CFPS data is fairly close to the census data, which is about 70% rural and 30% urban. Although the Chi-square test of the two-way contingency table is significant, but the Chi-square value is only 96.6. The frequency distribution of CHFS is also very similar to the census data, though statistically significantly different ($\chi2(1) \approx 150000$，$p = 0.000$).

Table 32. Rural-Urban Distribution (%)

	CFPS 2010			
	T1 members	Individual respondents	2010 census	CHFS 2010[b]
By National Bureau of Statistics				
Rural	55.4	54.5	49.7	48.7

Urban[a]	44.7	45.5	50.3	51.3
N	36,571	27,444	1,332,810,869	528,808,705
By Village/Neighborhood Community				
Village	69.6	68.4	—	—
Neighborhood community	30.4	31.6	—	—
N	36,571	27,444	—	—
By *hukou*				
Agricultural	—	73.6	70.9	63.3
Non-agricultural	—	26.4	29.1	36.7
N	—	27,204	131,9046,434	440,104,614

[a] "Towns" and "cities" of the census data were merged as "urban".
[b] The data of CHFS are weighted.

7.5 Educational Level

The distribution of the highest educational level of T1 members of the CFPS baseline survey is very close to the 2010 national census. The difference between these two data is mainly the slightly lower proportion of junior middle school graduates among the T1 members (a difference of about 5%). Apart from this, the differences in other educational levels are basically less than 2%. The Chi-square test of the two-way contingency table shows a significant difference ($\chi^2(6) = 493.1084$， $p = 0.000$). We took the group of "illiterate/semi-illiterate" as the reference group and the census data as the exposure. The log-rate analysis shows a significant difference in the proportions of "junior middle school", "senior middle school", "junior college" and "graduate".

Table 33. Highest Level of Education, Age 6 and above (%)

	CFPS 2010		2010 census	CGSS 2010
	T1 members	Individual respondents		
Illiterate/semi-illiterate	24.31	29.7	22.9	22.7
Primary school	21.54	23.6	19.9	19.0
Junior middle school	32.05	26.2	37.6	26.3
Senior middle school	14.35	13.2	11.8	18.3
Junior college	4.65	4.3	4.7	7.7
College	2.93	2.9	2.8	5.5
Graduate	0.2	0.2	0.3	0.6
Sample size (N)	29,974	23,219	111,601,269	10,173

The comparison between CFPS and CGSS shows that the respondents whose highest educational level is junior middle school have a higher proportion among the T1 members of CFPS than the CGSS, but the proportions are lower for the categories of "senior middle school",

"junior college", "college" and "graduate". We tested the differences to evaluate the quality of the CFPS data: for the comparison between CGSS and the census, the Chi-square test result of the two-way contingency table is significant, with a value of about 1200, much higher than the Chi-square value from comparing CFPS and the census; log-rate analysis shows that all the groups of CGSS are significantly different from the reference groups except for "primary school". Therefore, we suspect that the source of such differences may be CGSS' over-sampling of senior middle school graduates and insufficient representation of lower educational levels.

Compared with CFPS T1 data, the respondents of the individual questionnaires in CFPS have lower educational level, which might be due to the underestimation. Family members with lower level of education are highly likely to stay at home, thus this group of people are more likely to complete individual questionnaire while family members with higher level of education are more likely not presenting at the time of interview. In addition, it is possible that T1 respondants who answered questions on behalf of their family members intend to report a higher level of education rather than the actual situation.[61]

7.6 Marital Status

In addtion, we compared the distributions of marital status of the population aged 15 or above between CFPS, CHFS, CGSS, and the 2010 national census to test the quality of the marriage data of CFPS. Table 34 lists the results of the comparisons. The CFPS data again have two versions—the T1 family members and the respondents of the individual questionnaires. The CHFS asks marital status for all family members. The CGSS only includes marital status for respondents aged 18 or above. The statistical results show that the marital status distribution of T1 members of CFPS is very close to the census. The two distributions are essentially identical, although the huge sample size of the census leads to a significant difference in Chi-square test of the two-way contingency table ($\chi^2(3) = 23.5398$， $p = 0.000$), as well as in the log-rate analysis with the census data as the exposure (χ^2 (3) =24.15, p = 0.000) .

Table 34. Distribution of Marital Status, Ages 15 or above (%)

		CFPS 2010				
		T1 members	Individual respondents	2010 Census	CHFS 2011[a]	CGSS 2010[b]
Total	Unmarried	20.8	14.6	21.6	18.2	8.1
	Married	72.2	78.5	71.3	76.5	82.8
	Divorced	1.2	1.2	1.4	1.3	2.1
	Widowed	5.8	5.8	5.7	4.0	7.0
	N	30,642	22,197	105,542,243	24,693	10,154
Male	Unmarried	24.1	17.0	21.6	21.1	10.1
	Married	71.2	78.2	71.3	75.3	83.6
	Divorced	1.4	1.4	1.4	1.2	2.1
	Widowed	3.3	3.4	5.7	2.3	4.1

[61] Technical report: CFPS-21.

	N	15,454	10,732	52,943,450	12,352	4,932
Female	Unmarried	17.3	12.3	18.5	15.2	6.3
	Married	73.2	78.8	72.3	77.7	82.0
	Divorced	1.0	0.9	1.2	1.4	2.0
	Widowed	8.5	8.0	8.0	5.7	9.7
	N	15,188	11,465	52,598,793	12,341	5,222

[a] The data of CFPS 2010 are weighted.
[b] The data of CGSS 2010 represent the population at ages 18 or above.

As the proportion of youth among CFPS individual respondents is much lower than the census, there is a difference in the marital status distribution between CFPS individual respondents and the census data ($\chi^2(3) = 669.7521$， $p = 0.000$). The proportion of unmarried group is lower, while proportions of the divorced and the widowed are not different from the census. This indicates that the sample of CFPS individual respondents have only one source of bias—under-representation of youth.

The distribution of CHFS data is also quite similar to the census but the difference is larger than that between CFPS T1 members and the census ($\chi^2(3) = 157.7124$, p=0.000). Because CGSS focused on family members aged 18 or above and over-sampled of urban population, the marriage distribution is significantly different from other data.

The sex-specific marital status distributions of CFPS T1 family members are also consistent with the census data, with a higher proportion of unmarried males—but the female results are essentially the same.

8. Weights Calculations

The CFPS 2010 provided weights for family, adult, and child data sets for national full sample and national representative sample. The national full sample weight is the combined weights of five "large provinces" (Shanghai, Henan, Gansu, Liaoning, and Guangdong) and one "small province" (other provinces among 25 provinces/cities/autonomous regions). The national representative sample weight is the combined resample weights of the five "large provinces" and "small provinces". Given the basic weights design, we calculated weights at all three levels of data—communities, families, and individuals. Data users are provided with three types of weights: sampling design weights, non-response adjustment weights, post-hoc stratification adjustment weights, and trimming of the weights.

The s design weights are the inverses of the multiplying sampling probabilities from the first, second, and third stage sampling.[62] As for the resampled sample, the calculation also took into account the probabilities of resample from the sample districts/counties in the first stage.

Non-response adjustment weights adjust the non-responses at the family member and family/individual questionnaire levels. At the family member level, the non-response weight is based on weighted adjustment method, which used the number of completed family roster questionnaires over the total number of families of the community as the weighted adjustment coefficient. At the family/individual questionnaire level, the non-response weight is also based on weighted adjustment method. The adjustment coefficient is the number of family households completed the quesionnaires over the total number of family households expected to answer the survey. The non-response weight for the individual questionnaire use a logistic modeling estimating the likelihood of answering the quesionniare based on two stages as the adjustment coefficient. In the first stage, the individual sample was divided into contacted sample and non-contacted sample using supplemental information from the data to build the logistic regression model to obtain the probability of answering surveys at the indicidual connecting level. In the second stage, the connected sample is devided into refusal sample and non-refusal sample to build the logistic regression model and obtain the probability of refusing the interview.

The post-stratification aimed to reduce the sampling errors and increase the accuracy of the estimation by adjusting the structural biases of the sample due to the complexity of the sample design, the diversity in the field investigation, and the non-response. CFPS used gender, age, and rural-urban division to construct the post-stratification.

The trimmings of weights restricted the adjustment coefficients of non-response and post-stratification in a certain range to control the standard errors after the non-response and post-stratification adjustments. In addition, the trimming restricted the final weights after the sampling design weight, non-response adjustment, and post-stratification adjustment in a certain range to secure the efficiency of the estimation.

Finally, in order to equal the final weights and the population total, more adjustments were needed. After the adjustments, we finally get the full sample weights data sets of 25 provinces

[62] The third stage sampling seeks to adjust for multiple eligible households that live at the same sampled address. One household was randomly selected in such case.

and the national representative sample weights of the resample from the 25 provinces. Each of the data set contained weigths for family, adult, and children. The full sample weights datasets of 25 provinces are the combination of the weights data sets of “large provinces” and “small provinces.”

The use of weights is associated with data users’ target population and types of data sets. The detailed information regarding the different types of data sets and the represented total population can be found in Table 15. Table 35 lists the weighting variables in CFPS.

Table 35. Names and Labels of the Weighting Variables in CFPS

Data Set	Variable Name	Variable Label
Family	fswt_nat	Family weight-national full sample
	fswt_res	Family weight-national resample
Adult/Child	rswt_nat	Individual weight-national full sample
	rswt_res	Individual weight-national resample

For detailed calculation methods of the weights and the weighted analysis results of each CFPS data set, see *CFPS Baseline Survey Weights Calculation* (CFPS-17).9. Technical Reports

9. Technical Reports

In addition to this handbook, a series of technical reports would be released together with the CFPS 2010 data. The reports address different aspects of the CFPS project, and help users gain a better understanding of the survey and the data. The current list of the technical reports is provided below. More reports will be added in future work on the survey project.

CFPS-1: *Sampling Design*. Yu Xie, Zeqi Qiu, Ping Lv, 2012.

CFPS-2: *Third-stage Sampling Frame Construction*. Hua Ding, 2012.

CFPS-3: *The Implementation Report*. Hua Ding, 2012.

CFPS-4: *Quality Supervision Report*. Jie Yan, Yi Sun, Xueliang Teng, Liying Ren, Yan Sun, 2012.

CFPS-5: *Sample Contact Results*. Yan Sun, 2012.

CFPS-6: *Compilation and Data Cleaning of Family Relations Data Set*. Yuhuan Sun, Yu Xie, Jingwei Hu, Chunni Zhang, Qi Xu, Guoying Huang, 2012.

CFPS-7: *Data Cleaning of the Family Relations Data Set*. Qi Xu, Chunni Zhang, Yuhuan Sun, Jingwei Hu, Ping Lv, 2012.

CFPS-8: *Occupational and Industry Codes*. Liying Ren, Li Li, Chao Ma, 2012.

CFPS-9: *Career Expectation Codes*. Yu Xie, Wangyang Li, Chao Ma, Guoying Huang, Airan Liu, 2012.

CFPS-10: *Conversion of Occupational Codes and Construction of Socioeconomic Indices (CFPS-10).*. Guoying Huang, Yu Xie, 2012.

CFPS-11: *Composite Variables (I): Verbal and Mathematical Tests*. Hongwei Xu, Weixiang Luo, 2012.

CFPS-12: *Composite Variables (II): Educational Level and Depression Scale*. Yu Xie, Qi Xu, Chunni Zhang, Hongwei Xu, 2012.

CFPS-13: *Composite Variables (III): "Best Variables"for Age and Marriage*. Chunni Zhang, Qi Xu, Yan Sun, 2012.

CFPS-14: *Adjustments of Rural Family Income*. Yu Xie, Chunni Zhang, Guoying Huang, Qi Xu, Hongwei Xu, 2012.

CFPS-15: *Income and Expenditure Data*. Yan Shen, Xiaoyan Lei, 2012.

CFPS-16: *Income Comparisons between CGSS, CHIP, CHFS, and CFPS.* Qi Xu, Chunni Zhang, Xiang Zhou, Yu Xie, 2012.

CFPS-17: *Weight Calculation*. Ping Lv, Yu Xie, 2012.

CFPS-18: *Sample Maintenance*. Ping Lv, 2012.

CFPS-19: *Poverty Rate Comparisons between CFPS, CGSS, CHIP, and CHFS*. Chunni Zhang, Qi Xu, Xiang Zhou, Xiaobo Zhang, Yu Xie, 2012.

CFPS-20: *Dialect Codes*. Lingwei Wu, Wenjia Zhang, 2013

CFPS-21: *Evaluation and Analysis of Differences in Self-reporting and Proxy on Education Level.* Guoying Huang, Jingwei Hu, Yu Xie, Chunni Zhang, 2013

CFPS-22: *Composite Variables (IV): Parents' Social Status.* Chunni Zhang, Hua Ye, Lihong Dai, Jingwei Hu, Yu Xie, 2013

10. References

The Institute of Social Science Survey, Peking University, 2009. *China Report 2009*. Beijing: Peking University Press.

The Institute of Social Science Survey, Peking University, 2010. *China Report 2010*. Beijing: Peking University Press.

The Institute of Social Science Survey, Peking University, 2011. *China Report 2011*. Beijing: Peking University Press.

Ren, Qiang and Xie Yu. 2011. "Statistical Analysis of Longitudinal Data." *Population Research* 35(6): 3-12.

Sun, Yan, Yan Jie, Ding Hua, Gu Jiafeng, Liu Yue, Yao Jiahui and Zou Yanhui. 2011. *China Family Panel Studies 2010 – Interviewer Training Handbook*. Beijing: Peking University Press.

Xie, Yu. 2006. *Sociological Methodology and Quantitative Research*. Beijing: Social Sciences Academic Press (China).

Xie, Yu. 2010. "Understanding the Inequality in China." *Chinese Journal of Sociology* 30(3): 1-20.

Xie, Yu and Miranda Brown. 2009. "Between Heaven and Earth: Dual Accountability of Han Officials." *Chinese Journal of Sociology* 31(4): 1-28.

Xie, Yu. 2011. "Evidence-Based Research on China: A Historical Imperative." *Chinese Sociological Review* 44(1)：14-25.